Fully Offline Reinforcement Learning

Mattie Fellows, Clarisse Wibault,
Uljad Berdica, Johannes Forkel,
Michael A. Osborne, Jakob N. Foerster

Keywords: Offline RL, Bayesian RL, Model-Based ORL, Regret Analysis

Summary

Offline RL (ORL) promises safe and sample-efficient deployment but existing methods rely on undocumented online interactions for hyperparameter tuning and lack reliable fully offline estimates of initial online performance. We introduce SOReL, a fully offline Bayesian model-based RL method that learns a posterior over dynamics, estimates policy value via predictive uncertainty, and enables complete offline hyperparameter selection. We further propose TOReL, which extends this tuning framework to arbitrary model-free and model-based ORL algorithms. We provide a regret analysis showing that Bayesian offline RL achieves the minimax-optimal parametric rate under standard regularity conditions. Together, our methods establish a practical and theoretically grounded framework for fully offline RL.

Contribution(s)

1. In Section 5.1 we derive the posterior information loss (PIL), an information-theoretic metric that upper bounds true regret in offline Bayesian RL and characterises how regret decreases through the model's information rate. Since the PIL is computable from offline data alone, it provides a principled metric for fully offline hyperparameter tuning.
   Context: None
2. In Section 5.3 we carry out a frequentist regret analysis of offline model-based Bayesian RL, proving that its regret under the true data-generating distribution converges at the minimax-optimal parametric rate $\mathcal{O}\left(1/\sqrt{N}\right)$ under standard local asymptotic normality assumptions.
   Context: Research on Bayesian offline RL remains limited. To the authors' knowledge, only Chen et al. (2024) formulate offline model-based RL as a BAMDP, and no theoretical regret analysis currently exists.
3. In Section 6.1, we introduce SOReL, a fully offline RL framework that unifies policy learning, evaluation, and entirely offline hyperparameter tuning. Policy values are estimated via the Bayesian predictive median, and hyperparameters are selected using the PIL and the value estimate as principled criteria.
   Context: Developing a fully offline pipeline for offline RL remains a largely under-explored open problem (Jackson et al., 2025). Off policy evaluation (OPE) methods (Fu et al., 2021) estimate policy value offline but are evaluators rather than planners, and thus do not account for offline policy improvement. Moreover, they typically require online tuning of their own hyperparameters. The closest method to SOReL is Paine et al. (2020), which likewise relies on online tuning.
4. In Section 6.2 we introduce TOReL, which extends SOReL's hyperparameter tuning as a general method for fully offline tuning of model-based and model-free offline RL algorithms including state of the art methods like IQL (Kostrikov et al., 2021) and ReBRAC (Tarasov et al., 2023).
   Context: As contribution 3, but TOReL is designed for settings where offline value estimation is not required.
5. In Section 7 we empirically confirm SOReL's ability to approximate online value and TOReL's ability to tune hyperparameters of general ORL algorithms, both entirely offline.
   Context: SOReL's offline approximate of true online returns has an average absolute error that is less than a quarter of the absolute error of FQE (used by Paine et al. (2020)) and TOReL matches or outperforms OPE-based selection methods while eliminating the online sample cost required by Jackson et al. (2025)'s online UCB tuning algorithm.

Fully Offline Reinforcement Learning

Mattie Fellows[1,⋆], **Clarisse Wibault**[1,2,⋆],
Uljad Berdica[1], **Johannes Forkel**[1],
Michael A. Osborne[2], **Jakob N. Foerster**[1]
matthew.fellows@eng.ox.ac.uk, clarisse.wibault@eng.ox.ac.uk

[1]**Foerster Lab for AI Research (FLAIR), Department of Engineering Science, University of Oxford**
[2]**Machine Learning Research Group, Department of Engineering Science, University of Oxford**

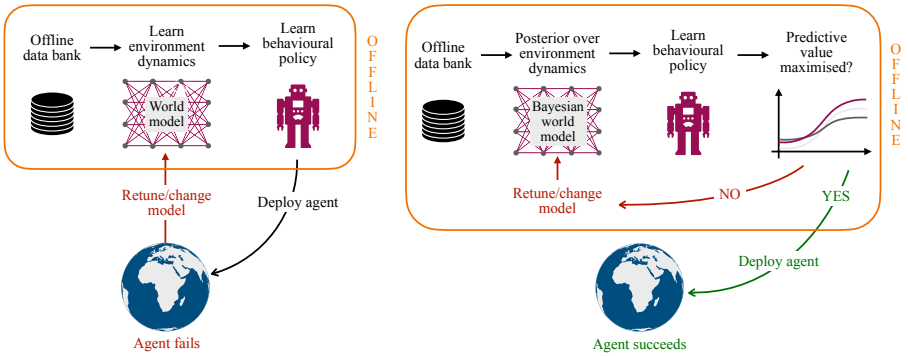
⋆ Equal Contribution

Abstract

Sample efficiency remains a major barrier to the real-world deployment of Reinforcement Learning (RL), where large numbers of online interactions are often costly or unsafe. Offline RL (ORL) seeks to address this challenge by learning policies from static datasets, yet existing methods rely on undocumented online interactions for hyperparameter tuning. Moreover, current methods provide no reliable estimate of their initial online performance using offline data alone. To overcome these limitations, we introduce two complementary algorithms for fully offline RL. Firstly, SOReL (Safe Offline Reinforcement Learning) is a model-based Bayesian approach that infers a posterior over environment dynamics from offline data and learns an approximation to a Bayes-optimal policy entirely offline. By quantifying predictive uncertainty via model ensembling, SOReL enables reliable and tractable offline prediction of online policy value and fully offline hyperparameter selection. Secondly, TOReL (Tuning for Offline Reinforcement Learning) extends this principle to arbitrary model-free and non-Bayesian ORL algorithms, leveraging predictive uncertainty quantification to eliminate costly online tuning loops. We additionally provide a regret analysis of offline Bayesian RL, showing that Bayesian approaches achieve the optimal frequentist minimax regret rate, thereby offering a formal frequentist justification for offline Bayesian methods. Empirical results demonstrate that SOReL accurately predicts online performance and that, using only offline data, TOReL achieves competitive results with online hyperparameter tuning methods, advancing safe and reliable fully offline RL for real-world deployment. Our implementations are publicly available.

1 Introduction

Offline RL (Lange et al., 2012; Levine et al., 2020; Murphy, 2024) aims to learn effective policies from static datasets, enabling agents to act autonomously and safely upon deployment. However, existing offline RL methods (Tarasov et al., 2023; Kostrikov et al., 2021; Kidambi et al., 2020; Yu et al., 2021) fall short of this promise due to two largely overlooked issues. **Issue I: there is no *fully offline* metric to tune all hyperparameters or select between approaches.** In practice, state-of-the-art methods rely on *online* interactions to perform extensive hyperparameter tuning (Zhang et al., 2021; Jackson et al., 2025). As illustrated in Fig. 1a, this leads to iterative cycles of offline training, online evaluation, hyperparameter adjustment, and retraining, incurring substantial online sample complexity — precisely what offline RL seeks to avoid. For performance-critical problem settings, we also need reliable online performance guarantees *before* the agent is deployed online. Precisely, **issue II: current methods offer no reliable *fully offline* method to approximate true *online* value**, i.e. the online expected returns of a policy trained using offline data. This is concerning from an AI



(a) Existing Model-Based Offline RL (b) SOReL and TOReL (Our Approach)

Figure 1: Existing ORL algorithms rely on online interactions for hyperparameter tuning and validation, leading to poor online sample efficiency. In contrast, SOReL and TOReL carry out fully offline hyperparameter adjustment, using the predictive value and PIL as a tuning signal. Only then is the agent trained and deployed.

safety perspective, as without a reliable performance estimate, we cannot deploy agents into the real world where agent failure presents a serious hazard to human life; for these settings it is essential that agents are deployed with near-optimal expected returns. For less safety-critical domains, users still require deployment-time guarantees, as suboptimal policies incur quantifiable costs in terms of reduced expected returns, such as financial losses or degraded operational performance.

We address both issues by leveraging model-based RL and exploiting predictive uncertainty from a posterior distribution over environment dynamics learned from offline data. We develop a fully offline RL pipeline (Fig. 1b) comprising two complementary algorithms. Firstly, for settings requiring accurate offline value prediction, we introduce **SOReL** (Safe Offline Reinforcement Learning), a theoretically grounded Bayesian model-based approach that resolves **Issues I & II**. The posterior and learned dynamics define a Bayesian RL problem whose solution, the Bayes-optimal policy, is robust to model uncertainty in environment dynamics when deployed online. For tractability, we approximate posterior sampling using ensembling and learn an amortised Bayes-optimal policy via meta-RL rollouts in posterior-sampled environments. The median predictive return across rollouts provides an offline estimate of online performance, while the predictive variance quantifies its reliability, directly addressing **Issue II**.

To enable fully offline hyperparameter tuning, we introduce the *posterior information loss* (PIL), defined as the expected KL divergence between the posterior model and the true dynamics. We show that the PIL upper-bounds regret and governs its convergence rate. Under standard regularity conditions, Bayesian offline RL achieves the optimal frequentist minimax rate, establishing a formal bridge between Bayesian uncertainty quantification and frequentist guarantees. Minimising and calibrating the PIL enables principled tuning of the environment dynamics model and inference hyperparameters, while the predictive median is used to tune planner hyperparameters, thereby resolving **Issue I**.

Empirically, we evaluate SOReL in a zero-shot setting on standard MuJoCo offline RL benchmarks (Yu et al., 2020; Kidambi et al., 2020). SOReL's predictive value closely tracks realised online returns, with true performance consistently contained within the posterior predictive range once the PIL is calibrated. Compared to an OPE-based approach that still requires online tuning (Paine et al., 2020), SOReL achieves less than one-quarter of the average absolute error, providing substantially more accurate value estimation, as well as calibrated uncertainty and conservative behaviour in low-data regimes. SOReL achieves this entirely without online interaction.

Secondly, for settings where exact offline value prediction is unnecessary, we introduce **TOReL** (Tuning for Offline Reinforcement Learning), which extends our fully offline tuning procedure to address **Issue I** in general non-Bayesian and model-free settings, enabling compatibility with existing methods. Many prior approaches (Yu et al., 2020; Kidambi et al., 2020; Ball et al., 2021; Lu et al., 2022b; Sun et al., 2023; Tarasov et al., 2023; Sims et al., 2024) rely on heuristics tailored to standard ORL benchmarks, which typically exhibit limited behavioural diversity and state-action coverage (Fu et al., 2020); as a result, these methods can outperform more general Bayesian approaches like

SOReL on such benchmarks. TOReL leverages the PIL and a predictive value metric correlated with true performance to tune hyperparameters entirely offline. Since both SOReL and TOReL rely only on ensembling, already standard in modern offline RL (Jackson et al., 2025), our approach introduces no additional computational overhead.

Empirically, TOReL matches or outperforms off-policy evaluation (OPE)-based tuning approaches that still rely on online calibration (Paine et al., 2020). Crucially, unlike OPE methods, TOReL selects hyperparameters entirely offline using the PIL signal and predictive value metric. Compared to a state-of-the-art online bandit-based tuning method (Jackson et al., 2025), TOReL achieves comparable or superior performance while eliminating the substantial online sample cost required for hyperparameter search.

2 Preliminaries

Let X be a $\mathcal{X} \subseteq \mathbb{R}^n$-valued random variable. We denote its distribution as P_X and its density (if it exists) as $p(x)$. We denote the set of all distributions over $\mathcal{X}$ as $\mathcal{P}(\mathcal{X})$. We introduce the notation $\mathcal{G}(p)$ to represent the geometric distribution and $\mathcal{AG}(p)$ to represent the arithmetico-geometric distribution, with probability mass functions: $P_{\mathcal{G}}(i) \coloneqq (1-p)p^i$ and $P_{\mathcal{AG}}(i) \coloneqq (1-p)^2 p^i(i+1)$ respectively, for $i \in \mathbb{N}_0$ and parameter $p \in [0, 1)$. We denote the uniform distribution over $\{0, 1, \dots i\}$ as $\mathcal{U}_i$ and the multivariate normal distribution with mean vector μ and covariance matrix Σ as $\mathcal{N}(\mu, \Sigma)$.

2.1 Offline Reinforcement Learning

In our Reinforcement Learning (RL) setting, an agent is tasked with solving the learning problem in an infinite-horizon, discounted Markov decision process (Bellman, 1956; 1958; Puterman, 1994; Sutton & Barto, 2018): $\mathcal{M}^\star \coloneqq \langle \mathcal{S}, \mathcal{A}, P_0, P_S^\star(s,a), P_R^\star(s,a), \gamma \rangle$, with state space $\mathcal{S}$, action space $\mathcal{A}$ and discount factor γ. At time $t = 0$, an agent starts in an initial state allocated according the the initial state distribution: $s_0 \sim P_0$. At every timestep t, an agent in state s_t takes an action according to a policy $a_t \sim \pi(h_t)$, receives a scalar reward $r_t \sim P_R^\star(s_t, a_t)$ and transitions to a new state $s_{t+1} \sim P_S^\star(s_t, a_t)$ where $h_t \coloneqq \{s_0, a_0, r_0, s_1, a_1, r_1, \dots a_{t-1}, r_{t-1}, s_t\} \in \mathcal{H}_t$ is the observed a history of interactions with the environment. Here $\mathcal{H}_t \coloneqq \mathcal{S} \times (\mathcal{A} \times \mathbb{R} \times \mathcal{S})^t$ denotes the corresponding product space. We assume rewards are bounded with $r_t \in [r_{\min}, r_{\max}] \subset \mathbb{R}$ where $r_{\min}$ and $r_{\max}$ denote the minimum and maximum reward values respectively. For convenience, we often write the joint state transition-reward distribution as $P_{R,S}^\star(s,a)$. We denote the distribution over history h_t as $P_{t,\pi}^\star$. The goal of an agent is to learn an optimal policy $\pi^\star \in \Pi^\star$ where $\Pi^\star \coloneqq \arg\max_\pi J^\pi(\mathcal{M}^\star)$ is the set of policies that maximise the expected discounted return:

$$J^\pi(\mathcal{M}^\star) \coloneqq \mathbb{E}_{h_\infty \sim P_{\infty,\pi}^\star} \left[\sum_{i=0}^{\infty} \gamma^i r_i \right].$$

If the true transition and reward distributions $P_S^\star(s,a)$ and $P_R^\star(s,a)$ are known a priori, the problem reduces to planning, and (given sufficient compute) an optimal policy can be computed without further environment interaction; in this case, policies need only condition on the current state s_t. In contrast, we consider the learning setting where the dynamics are unknown. Under our Bayesian formulation, policies condition on the full history h_t, which carries the information required to update beliefs over the environment.

In offline RL (Lange et al., 2012; Levine et al., 2020; Murphy, 2024), an agent has access to a dataset of histories of various lengths collected from the true environment. The policies used to collect the data may vary and not be optimal. In this paper, we consider a zero-shot setting where the agent has no ability to interact with the target environment before deployment, as formalised by Jackson et al. (2025). The agent is then deployed at test time $t = 0$ and its performance evaluated. The goal of offline RL is to take advantage of offline data so that the deployed policy will be near-optimal from the outset. In model-based offline RL, a model of the environment dynamics is learned using the dataset. An offline policy is obtained by solving a planning problem using the learned dynamics model. In contrast, model-free offline RL methods forgo explicitly modelling environment dynamics and instead learn a policy or value function directly from the fixed dataset, without constructing an explicit dynamics model.

3 Related Work

3.1 Off-Policy Evaluation

Off-policy evaluation (OPE) algorithms (Le et al., 2019; Nachum et al., 2019; Thomas & Brunskill, 2016; Kostrikov & Nachum, 2020; Fu et al., 2021) estimate the performance of a target policy using data collected by a different behaviour policy. In principle, OPE provides an offline mechanism for approximating online value. However, OPE methods are evaluators rather than planners: they estimate the value of a fixed policy without performing policy improvement (Uehara et al., 2022) and therefore do not constitute a complete offline RL pipeline. Although many OPE methods attach uncertainty measures or confidence intervals, these are typically conditional on modelling assumptions and do not provide a principled diagnostic for when insufficient data, support mismatch, or model misspecification render estimates unreliable. Moreover, OPE methods generally require calibration of their own hyperparameters using online data and methods based on importance sampling (Kostrikov et al., 2021) and double robustness (Thomas & Brunskill, 2016) require datasets containing trajectories rather than individual datapoints.

OPE can also be used for offline hyperparameter selection by ranking candidate policies according to their estimated value. To our knowledge, Paine et al. (2020) provide the primary study of this approach using FQE. However, their results show substantial value overestimation in several domains, provide no principled signal of estimate reliability, and require online tuning of FQE hyperparameters, so the overall pipeline is not fully offline. As noted by Jackson et al. (2025), this framework is further limited to behavioural cloning and two critic-based methods that have since been surpassed by modern offline RL algorithms. Consequently, while OPE addresses offline value estimation, it does not fully resolve **Issues I & II**: both policy-learning and OPE-specific hyperparameters typically require calibration beyond the offline dataset, and existing approaches lack a principled mechanism to determine whether the available data suffice for reliable performance estimation. In contrast, our method derives an information-theoretic signal (the PIL) tied to regret that governs both hyperparameter tuning and predictive value reliability entirely offline

3.2 Model-Based Offline RL

Developing reliable model-based offline RL remains challenging. Beyond **Issues I & II**, performance depends critically on accurate dynamics modelling, as model errors can compound over long horizons. Additionally, benchmark datasets often lack coverage in important regions of state-action space, creating further generalisation challenges. Most existing approaches focus on mitigating this latter issue by introducing reward pessimism or uncertainty-based regularisation (Yu et al., 2020; Kidambi et al., 2020; Kumar et al., 2020; Yu et al., 2021; Fujimoto & Gu, 2021; Kostrikov et al., 2021; An et al., 2021; Ball et al., 2021; Lu et al., 2022b; Sun et al., 2023; Tarasov et al., 2023; Sims et al., 2024). Unifloral (Jackson et al., 2025) unifies these approaches within a common framework with streamlined implementations and benchmarking protocols, which we follow in our experiments.

Limited prior work directly addresses **Issues I & II**. Smith et al. (2023) propose offline hyperparameter tuning, but are limited to the model-free imitation learning setting and offer no value estimation. Wang et al. (2022) introduce offline pre-selection of hyperparameters for online methods, but do not learn optimal policies fully offline or approximate value. To our knowledge, our method is the first fully offline RL framework that integrates policy learning, value estimation (when required), and complete hyperparameter tuning using only offline data. Finally, Bayesian perspectives on offline RL remain underexplored. Chen et al. (2024) formulate offline model-based RL as solving a BAMDP, however no regret analysis of the Bayes-optimal policy is carried out, a continuous BAMDP (Guez et al., 2014) approximation is used to learn behavioural policies, does not propose a method for offline policy value estimation, and the algorithm still relies on online data for tuning.

4 Bayesian Offline RL

We now introduce our Bayesian RL (BRL) framework, which constitutes learning an (approximate) posterior over environment dynamics from offline data, and then solving a BRL problem to find a Bayes-optimal policy with that posterior acting as a prior over the space of MDPs. Bayesian RL

methods naturally minimise Bayes-regret, which maximises the average performance over MDPs as weighted by the posterior. We provide an introductory primer on Bayesian RL in Section A.

4.1 Learning a Posterior with Offline Data

A Bayesian epistemology characterises the agent's uncertainty in the MDP through distributions over any unknown variable (Martin, 1967; Duff, 2002). For our model-based approach, we characterise uncertainty in the environment dynamics. We first specify a parametric model of the unknown joint reward-state transition distribution: $P_{R,S}(s_t, a_t, \theta)$, with each $\theta \in \Theta \subseteq \mathbb{R}^d$ representing a hypothesis about the underlying true MDP $\mathcal{M}^\star$. A prior distribution over the parameter space P_Θ is specified, which represents the initial *a priori* belief over hypothesis MDPs before the agent has observed any transitions. As we show in Section 5, our results can easily be generalised to non-parametric methods like Gaussian process regression (Rasmussen & Williams, 2006; Wiener, 1923; Krige, 1951). We denote an offline dataset of N state-action-state-reward transition observations as: $\mathcal{D}_N = \{(s_i, a_i, s'_i, r_i)\}_{i=0}^{N-1}$, all collected from a single MDP $\mathcal{M}^\star$. Datapoints may be collected from several policies and non-Markovian sampling. Given the dataset $\mathcal{D}_N$, the prior P_Θ with density $p(\theta)$ is updated to posterior $P_\Theta(\mathcal{D}_N)$ with density $p(\theta|\mathcal{D}_N)$, using Bayes' rule:

$$p(\theta|\mathcal{D}_N) = \frac{p(\mathcal{D}_N|\theta)p(\theta)}{p(\mathcal{D}_N)} = \frac{\prod_{i=0}^{N-1} p(r_i, s'_i|s_i, a_i, \theta)p(\theta)}{\int_\Theta \prod_{i=0}^{N-1} p(r_i, s'_i|s_i, a_i, \theta)p(\theta)d\theta}. \tag{1}$$

The posterior represents the agent's updated belief in the unknown environment dynamics once $\mathcal{D}_N$ has been observed. We now detail how a Bayes-optimal policy is learned using the posterior as the initial belief in the environment dynamics.

4.2 Approximating a Bayes-Optimal Policy

Solving a BRL problem exactly is intractable for all but the simplest models (Martin, 1967; Duff, 2002; Guez et al., 2012; 2013; Zintgraf et al., 2020; Fellows et al., 2024), and exact posterior inference in Eq. (1) is typically infeasible for dynamics models of interest (e.g., nonlinear Gaussian world models). We therefore learn an approximate posterior $\hat{P}_\Theta(\mathcal{D}_N) \approx P_\Theta(\mathcal{D}_N)$ using randomised prior (RP) ensembling (Osband & Van Roy, 2017; Osband et al., 2018; Ciosek et al., 2020) (details in Section F.2). In the ORL setting, the prior P_Θ is replaced with the informative posterior $P_\Theta(\mathcal{D}_N)$, reducing the hypothesis space of the BRL problem. A Bayes-optimal policy is then learned using the posterior as the initial belief over environment dynamics, yielding the offline Bayesian RL objective:

$$J^\pi_{\text{Bayes}}(P_\Theta(\mathcal{D}_N)) \coloneqq \mathbb{E}_{\theta \sim P_\Theta(\mathcal{D}_N)} \left[\mathbb{E}_{h_\infty \sim P^\pi_\infty(\theta)} \left[\sum_{i=0}^{\infty} \gamma^i r_i \right] \right]. \tag{2}$$

Solving Eq. (2) for π to find a Bayes-optimal policy $\pi^\star_{\text{Bayes}}(h_t) \in \arg\max_\pi J^\pi_{\text{Bayes}}(P_\Theta(\mathcal{D}_N))$ can be formulated as a meta-RL problem (Zintgraf et al., 2020; Beck et al., 2024), enabling an RL^2-style procedure (Duan et al., 1987): environments are sampled from the posterior and the policy is trained via rollouts in these sampled models. An approximate Bayes-optimal policy is thus obtained by searching over history-conditioned policies, for example using recurrent PPO (Schulman et al., 2017). Crucially, posterior reasoning is amortised into the policy during training; deployment requires only a single forward pass of the recurrent network, incurring no additional overhead relative to standard model-based offline RL. Implementation details are provided in Section F.

Beyond enabling exploration when deployed, the Bayesian formulation provides explicit epistemic uncertainty in the learned dynamics model, which underpins our fully offline pipeline. To address **Issue II**, we estimate policy value using posterior predictive rollouts: the median return across sampled models yields a Bayesian estimate of test-time performance, while the predictive variance quantifies its uncertainty. To address **Issue I**, this predictive estimate enables fully offline tuning of planner hyperparameters. In addition, as shown in Section 5.1, the Posterior Information Loss (PIL) links posterior uncertainty to regret and is used to tune model and inference hyperparameters. Together, the predictive variance and PIL provide principled diagnostics of model misspecification, data coverage, and uncertainty calibration prior to deployment.

Finally, we remark that a Bayesian approach is relatively simple and general compared to existing model-based approaches in Section 3 as it does not rely on hand-crafted heuristics tailored to specific problem settings. We emphasize that adopting a Bayesian perspective does not introduce additional computational overhead beyond what is already standard in modern model-based offline RL. We rely only on ensembling for approximate posterior sampling — already standard in modern model-based offline RL (Jackson et al., 2025). No expensive posterior inference or additional planning modules are required; the predictive median, variance, and PIL are computed from the same ensemble rollouts used for policy learning, effectively extracting additional information from computations that are already required. Hence our method does not introduce new costly components, but instead leverages existing tools more efficiently to obtain principled tuning signals and reliable offline value estimates.

5 Regret Analysis

In this section, we provide a formal theoretical regret analysis of Bayesian offline RL. Our first result in Section 5.1 proves that regret is tied to the posterior information loss (PIL), which can be used as a tuning metric and diagnostic tool during offline RL training. In Section 5.3, we then use this result to prove that the Bayes-optimal policy convergences to zero regret at the minimax optimal rate under standard assumptions, thereby providing a formal frequentist justification for Bayesian offline RL in general. Proofs for all theorems can be found in Section D.

5.1 Derivation of Posterior Information Loss (PIL)

Given a dataset of datapoints N collected from the true MDP $\mathcal{M}^\star$, we can measure how far the performance of the Bayes-optimal policy is from the true optimal policy using the *true regret*, which, in this case, is the difference between the expected return $J^{\pi^\star}(\mathcal{M}^\star)$ of an optimal policy for the true MDP and the expected return of the of the Bayes-optimal policy $J^{\pi^\star_{\text{Bayes}}}(\mathcal{M}^\star, \mathcal{D}_N)$ given a posterior $P_\Theta(\mathcal{D}_N)$, all in the true MDP $\mathcal{M}^\star$:

$$\text{Regret}(\mathcal{M}^\star, \mathcal{D}_N) \coloneqq J^{\pi^\star}(\mathcal{M}^\star) - J^{\pi^\star_{\text{Bayes}}}(\mathcal{M}^\star, \mathcal{D}_N).$$

Our first result shows that true regret can be bounded using the PIL, defined as:

$$\mathcal{I}^\pi_N \coloneqq \mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\mathbb{E}_{s,a\sim\rho^\star_\pi}\left[\text{KL}\left(P^\star_{R,S}(s,a)\|P_{R,S}(s,a,\theta)\right)\right]\right],$$

where $\rho^\star_\pi \coloneqq \mathbb{E}_{i\sim\mathcal{AG}(\gamma)}\left[\mathbb{E}_{j\sim\mathcal{U}_i}\left[P^\star_{j,\pi}\right]\right]$ is the arithmetico-geometric ergodic state-action distribution induced by a policy π and recall $P^\star_{j,\pi}$ denotes the distribution over a j-length trajectory h_j. $\rho^\star_\pi$ places mass over state-action pairs according to how much errors in the model influence the regret at each state. Regions of the state-action space that require more timesteps to reach from initial states are weighted significantly less than those that are encountered earlier and more frequently, as state errors encountered early accumulate in each prediction from that timestep onwards.

The PIL has an intuitive information-geometric interpretation: the inner expectation $\mathbb{E}_{s,a\sim\rho^\star_\pi}\left[\text{KL}\left(P^\star_{R,S}(s,a)\|P_{R,S}(s,a,\theta)\right)\right]$ measures the distance between the model and the true distribution in terms of the information lost when approximating $P^\star_{R,S}(s,a)$ with $P_{R,S}(s,a,\theta)$, averaged across all states. The PIL thus measures how close the posterior's belief is to the truth according to the average information lost under the posterior expectation. We now prove that regret is upper bounded by the PIL:

Theorem 1. *Let $\mathcal{R}_{\max} \coloneqq \frac{(r_{max}-r_{min})}{1-\gamma}$ denote the maximum possible regret for the MDP. Using the PIL: $\mathcal{I}^\pi_N$, the true regret is bounded as:*

$$\text{Regret}(\mathcal{M}^\star, \mathcal{D}_N) \leq 2\mathcal{R}_{\max} \cdot \sup_\pi \sqrt{1 - \exp\left(-\frac{\mathcal{I}^\pi_N}{1-\gamma}\right)}.$$

The key insight from Theorem 1 is that the rate at which regret decreases with N is governed by the rate at which the PIL decreases, which is known as the *information rate*. This measures how much information the model has gained from an incremental amount of data. Fast information rates imply that highly informative posteriors can be learned from limited data, as regret decreases at least as quickly. How fast the PIL decreases depends on the model specification, the prior used, the data

coverage and the underlying MDP. This makes it an ideal metric to hyperparameter associated with the dynamics model and approximate inference method.

5.2 Gaussian World Models

Formulating our bound in terms of the PIL ties the regret to the KL divergence over the reward-state model: $\mathrm{KL}\left(P^{\star}_{R,S}(s,a) \| P_{R,S}(s,a,\mathcal{D}_N)\right)$. Not only is this mathematically more convenient, but it means that the PIL is easy to estimate in practice. As an example, many methods use an ensemble of Gaussian reward and state transition models of the form introduced by Chua et al. (2018), where the individual reward and state transition models are given by:

$$P_R(s,a,\theta) = \mathcal{N}(r_\theta(s,a), \sigma_r^2(s,a)), \quad P_S(s,a,\theta) = \mathcal{N}(s'_\theta(s,a), I\sigma_s^2(s,a)), \tag{3}$$

with variances characterised by σ_r^2 and σ_s^2, mean reward function $r_\theta(s,a)$ and mean state transition function $s'_\theta(s,a)$. Using an ensemble of Gaussian World Models admits a natural separation of aleatoric and epistemic uncertainty: the aleatoric uncertainty can be represented by the individual output variances σ_r^2 and $I\sigma_s^2(s,a)$, or mean of the learned variances, and the epistemic uncertainty by the ensemble disagreement, or variance of the means. Using a Gaussian world model, we find the PIL takes a convenient and intuitive form. Let $r(s,a,\mathcal{D}_N) \coloneqq \mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[r_\theta(s,a)\right]$ and $s'(s,a,\mathcal{D}_N) \coloneqq \mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[s_\theta(s,a)\right]$ denote the Bayesian mean reward and state transition functions and $r^\star(s,a)$ and $s^{\star\prime}(s,a)$ denote the true mean functions. We define the mean squared error between the true and Bayesian mean functions as:

$$\mathcal{E}(\mathcal{D}_N, \mathcal{M}^\star) \coloneqq \mathbb{E}_{(s,a)\sim\rho^\star_\pi}\left[\frac{\|r(s,a,\mathcal{D}_N) - r^\star(s,a)\|_2^2}{2\sigma_r^2(s,a)} + \frac{\|s'(s,a,\mathcal{D}_N) - s^{\star\prime}(s,a)\|_2^2}{2\sigma_s^2(s,a)}\right], \tag{4}$$

and the predictive variance as:

$$\mathcal{V}(\mathcal{D}_N) \coloneqq \mathbb{E}_{(s,a)\sim\rho^\star_\pi}\left[\mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\frac{\|r(s,a,\mathcal{D}_N) - r_\theta(s,a)\|_2^2}{2\sigma_r^2(s,a)} + \frac{\|s'(s,a,\mathcal{D}_N) - s'_\theta(s,a)\|_2^2}{2\sigma_s^2(s,a)}\right]\right]. \tag{5}$$

We now re-write the PIL for the Gaussian world model using these two terms:

Proposition 1. *Using the Gaussian world model in Eq. (3), it follows that*

$$\mathcal{I}^\pi_N = \mathcal{E}(\mathcal{D}_N, \mathcal{M}^\star) + \mathcal{V}(\mathcal{D}_N).$$

Proposition 1 shows that the PIL is governed by i) the MSE of the point estimate $\mathcal{E}(\mathcal{D}_N, \mathcal{M}^\star)$, which characterises how quickly the Bayesian mean function converges to the true function; and ii) the predictive variance $\mathcal{V}(\mathcal{D}_N)$, which characterises the epistemic uncertainty in the model and is a purely Bayesian term. The PIL can easily be estimated by estimating $\mathcal{E}(\mathcal{D}_N, \mathcal{M}^\star)$ using the empirical MSE under the offline data distribution and estimating $\mathcal{V}(\mathcal{D}_N)$ using approximate posterior sampling. This decomposition provides a principled diagnostic mechanism based on the relative magnitudes and contraction behaviour of $\mathcal{E}(\mathcal{D}_N, \mathcal{M}^\star)$ and $\mathcal{V}(\mathcal{D}_N)$. In particular, imbalances between these terms can signal model misspecification, insufficient data coverage, or miscalibrated uncertainty. We exploit this property directly for hyperparameter tuning and model validation in Section 6.1.

5.3 Frequentist Justification of Bayesian Offline RL

As discussed in Section 4 and derived in Section A, all solutions to a Bayesian offline RL problem minimise Bayes regret, which provides a Bayesian justification for our method. We now provide a complementary frequentist justification. Specifically, we prove that the regret under the true data-generating distribution converges at the same $\mathcal{O}\left(1/\sqrt{N}\right)$ rate established in prior offline RL work (Yu et al., 2020; Kidambi et al., 2020) under standard local asymptotic normality (LAN) assumptions (Le Cam, 1953; Barron, 1988; Clarke & Barron, 1990; Komaki, 1996; Hartigan, 1998; Barron, 1999; Aslan, 2006). These assumptions are formally stated and discussed in Section D.3.

Frequentist analyses of offline RL span a broad spectrum of structural assumptions. At one extreme, minimax analyses evaluate performance uniformly over large classes of MDPs, effectively calibrating guarantees against the most challenging admissible environments (Wang et al., 2020). While such

worst-case results sharply delineate statistical limits, they are agnostic to how representative those environments are in practice. Consequently, under minimal structure one obtains intrinsic hardness results (Wang et al., 2020), whereas under additional structural or parametric assumptions, classical statistical efficiency becomes attainable (Agarwal et al., 2021; Li et al., 2022).

The $\mathcal{O}\left(1/\sqrt{N}\right)$ scaling we obtain is the canonical parametric asymptotic convergence rate and is known to be minimax-optimal in regular finite-dimensional models for offline RL (Agarwal et al., 2021; Li et al., 2022). Modern model-based offline RL methods operate precisely in such structured regimes, positing smooth parametric dynamics models and relying on likelihood-based estimation (Yu et al., 2020; Kidambi et al., 2020). From a Bayesian standpoint, LAN formalises this modelling commitment: a smooth, identifiable parametric family in which posterior uncertainty contracts at the canonical $\mathcal{O}\left(1/\sqrt{N}\right)$ parametric rate. Consequently, under the conditions of Assumption 1, our bound is asymptotically rate-tight within this nonlinear parametric function class.

Theorem 2. *Let the data be drawn from the underlying true distribution $\mathcal{D}_N \sim P^\star_{Data}$. Under standard local asymptotic normality assumptions (see Assumption 1 in Section D.3), there exists some constant $0 < C < \infty$ such that for sufficiently large N:*

$$\mathbb{E}_{\mathcal{D}_N \sim P^\star_{Data}}\left[\text{Regret}(\mathcal{M}^\star, \mathcal{D}_N)\right] \leq 2\mathcal{R}_{\max} \cdot \sqrt{1 - \exp\left(-\frac{Cd}{(1-\gamma)N}\right)} = \mathcal{O}\left(\sqrt{\frac{d}{(1-\gamma)N}}\right) \quad (6)$$

We also extend this analysis to include the case of model-misspecification and sub-optimal policy learning in Theorem 3, obtaining tighter bounds than obtained previous in literature (Kidambi et al., 2020) as $\gamma \to 1$. Finally, we remark that asymptotic frequentist regret bounds like Eq. (6) are primarily rate characterisations: they describe how regret scales in the large-sample limit but do not provide calibrated, dataset-specific diagnostics. Bounds of the form $\text{Regret}_N \leq C/\sqrt{N} + m$ depend on problem-specific constants C, m and unspecified large-sample regimes that are inaccessible in practice. Consequently, while such results are valuable for theoretical comparison, they offer limited use as practical metrics for assessing dataset sufficiency, posterior calibration, or hyperparameter selection in finite-sample settings. By contrast, the PIL metric yields directly observable quantities that can be monitored and acted upon during training.

6 Methods

6.1 SOReL

We now introduce SOReL in Algorithm 1, our fully offline Bayesian offline RL algorithm with value estimation and hyperparameter tuning. The algorithm shows a single step in an iterative process that updates the policy π and hyperparameters ϕ whilst returning a value estimate after each iteration and a set of metrics that include the PIL values and predictive variance. The metrics can be used for diagnostics and assessing when the desired value has been achieved depending upon the setting. In our SOReL framework, there are three sets of hyperparameters: ϕ_I the model (such as the architecture and hyperparameters for learning a neural-network function approximator); ϕ_{II} the approximate inference method (such as the number of ensemble members for RP); and ϕ_{III} the BAMDP solver (the hyperparameters of a Bayesian meta-learning algorithm like RNN-PPO). Sets ϕ_I and ϕ_{II} are tuned jointly to both minimise the PIL and ensure a roughly even split between the predictive variance and MSE loss terms. Set ϕ_{III} is then tuned to minimise approximate regret based on the now-fixed model and approximate posterior: for each combination of hyperparameters, we update the policy using the approximate BAMDP solver (BayesPolicyImprovement(·)), and choose the combination whose policy leads to the highest approximate value.

Algorithm 1 SOReL$(P_\Theta, \mathcal{D}_N, \phi, \pi)$

Hyperparameter tuning:
$\phi_I, \phi_{II} \leftarrow \arg\min_{\phi_I, \phi_{II}} \text{PIL}(\phi_I, \phi_{II}, \mathcal{D}_N)$
  s.t. $\mathcal{E}(\mathcal{D}_N, \mathcal{M}^\star) \approx \mathcal{V}(\mathcal{D}_N)$
$\phi_{III} \leftarrow \arg\max_{\phi_{III}} \text{PredValue}(\phi_I, \phi_{II}, \phi_{III}, \mathcal{D}_N, \pi)$
Policy Learning and Value Estimation:
$\pi \leftarrow \text{BayesPolicyImprovement}(\phi_I, \phi_{II}, \phi_{III}, \mathcal{D}_N)$
$J_N \leftarrow \text{PredValue}(\phi_I, \phi_{II}, \phi_{III}, \mathcal{D}_N, \pi)$
metrics $\leftarrow \text{GetMetrics}(\phi_I, \phi_{II}, \phi_{III}, \mathcal{D}_N, \pi)$
return ϕ, π, J_N, metrics

Fixing Issue I - PIL Monitoring: To tune sets ϕ_I and ϕ_{II}, we monitor the change in the PIL. Our objective is to select hyperparameters that minimise the PIL whilst ensuring that the MSE term (c.f. Eq. (4)) closely matches the predictive variance term (c.f. Eq. (5)), i.e. $\mathcal{E}(\mathcal{D}_N, \mathcal{M}^\star) \approx \mathcal{V}(\mathcal{D}_N)$. The

PIL alignment provides a powerful bias–variance diagnostic during training: $\mathcal{E}$ measures systematic model error (bias), whereas $\mathcal{V}$ captures epistemic uncertainty (posterior variance). If $\mathcal{E} \gg \mathcal{V}$ and fails to contract as N increases, the posterior is confident but biased, suggesting overfitting due to under-dispersed posterior uncertainty. Under a well-specified and well-calibrated model, both terms contract at comparable rates as N increases; persistent imbalance signals hyperparameter misconfiguration in ϕ_I and/or ϕ_{II}. Since miscalibrated uncertainty leads to unreliable value estimates, we tune ϕ_I and ϕ_{II} first in Algorithm 1. Empirically, as shown in Section 7.1, when $\mathcal{E} \approx \mathcal{V}$ and both are small, the approximate value closely tracks the true value. Likewise, in regimes with limited data coverage, policies must be regularised to remain close to the support of the training distribution (e.g. via a pessimistic prior in SOReL, or pessimistic mechanisms such as MOPO's reward penalty and MOReL's uncertainty-aware dynamics in TOReL). In this regime, value estimates are inherently biased, which is unavoidable in offline RL approaches with limited data coverage. However, this failure mode is reflected in an increase in ensemble variance due to the large epistemic uncertainty over regions with limited data, and consequently in the variance term of the PIL, which provides a practical signal that the policy is operating in poorly supported regions. We note that standard OPE methods do not expose this uncertainty.

Fixing Issues I and II - Bayesian Value Estimation: We estimate the value of the Bayes-optimal policy using the posterior predictive median:

$$\hat{J}^{\pi^\star_{\text{Bayes}}}(\mathcal{M}^\star) = \mathbb{M}_{\theta\sim P_\Theta(\mathcal{D}_N), h_\infty \sim P^\pi_\infty(\theta)}\left[R(h_\infty)\right], \tag{7}$$

where $\mathbb{M}_{\theta\sim P_\Theta(\mathcal{D}_N), h_\infty \sim P^\pi_\infty(\theta)}\left[R(h_\infty)\right]$ denotes the median predictive return. In Algorithm 1, $\hat{J}^{\pi^\star_{\text{Bayes}}}(\mathcal{M}^\star)$ is estimated in PredValue$(\cdot)$ via Monte Carlo sampling: ensemble members are drawn from the approximate posterior, the Bayes-optimal policy is rolled out in each sampled model, and the empirical median of returns is computed. Alternative Bayesian estimators, such as the predictive mean or lower quantiles, are also valid and correspond to different loss functions or levels of conservatism (see Section C). We adopt the predictive median because it is the Bayes estimator under absolute error loss and is robust to skewed or heavy-tailed predictive return distributions. Under symmetric predictive distributions the median and mean coincide, but in skewed posterior predictive settings the median provides greater stability. In particular, it avoids being overly conservative while remaining insensitive to outlier posterior samples corresponding to rare but highly optimistic or pessimistic dynamics. Our empirical evaluations in Section 7.1 support this choice. The posterior predictive median further enables fully offline tuning of hyperparameter set ϕ_{III}, by selecting the policy configuration that maximises the predictive value in Algorithm 1. Moreover, the predictive variance across rollouts characterises the epistemic and aleatoric uncertainty in the value estimate (returned in the 'metrics' set in Algorithm 1) and provides a quantitative signal for the practitioner to assess whether the learned policy is sufficiently reliable for deployment.

6.2 TOReL

Algorithm 2 TOReL$(P_\Theta, \mathcal{D}_N, \phi)$

$\phi_I, \phi_{II} \leftarrow \arg\min_{\phi_I,\phi_{II}} \text{PIL}(\phi_I, \phi_{II}, \mathcal{D}_N)$
$\quad$ [s.t. $\mathcal{E}(\mathcal{D}_N, \mathcal{M}^\star) \approx \mathcal{V}(\mathcal{D}_N)$, (model-based)]
$\phi_{III} \leftarrow \arg\max_{\phi_{III}} \text{ValueMetric}(\phi_I, \phi_{II}, \phi_{III}, \mathcal{D}_N)$
$\pi^\star_{\text{TOReL}} \leftarrow \begin{cases} \text{ORL}(\phi_{III}, \mathcal{D}_N), & \text{(model-free)} \\ \text{ORL}(\phi_I, \phi_{II}, \phi_{III}, \mathcal{D}_N), & \text{(model-based)} \end{cases}$
return ϕ, $\pi^\star_{\text{TOReL}}$

SOReL's offline hyperparameter tuning methods are directly applicable to general offline RL approaches, allowing us to address **Issue I** for existing offline methods. We now adapt these methods to derive a general tuning for offline reinforcement learning approach called TOReL, shown in Algorithm 2. A policy is learned offline using a planning algorithm, denoted by ORL. There thus exists a corresponding set of hyperparameters associated with ϕ_{III} the offline planner. For model-based methods with uncertainty estimation like MOReL Kidambi et al. (2020) and MOPO Yu et al. (2020), we can exactly adapt SOREL's PIL tuning method to the parameters associated with: ϕ_I the dynamics model and ϕ_{II} uncertainty estimation. For all other methods, we introduce and learn a dynamics model and an approximate inference method like in SOReL and jointly tune the corresponding hyperparameters ϕ_I and ϕ_{II} to minimise the PIL. Since the policy learned with ORL is typically neither Bayes-optimal nor robust to model uncertainty, we expect that applying SOReL's value estimation method to more general methods in TOReL will not yield an accurate estimate of

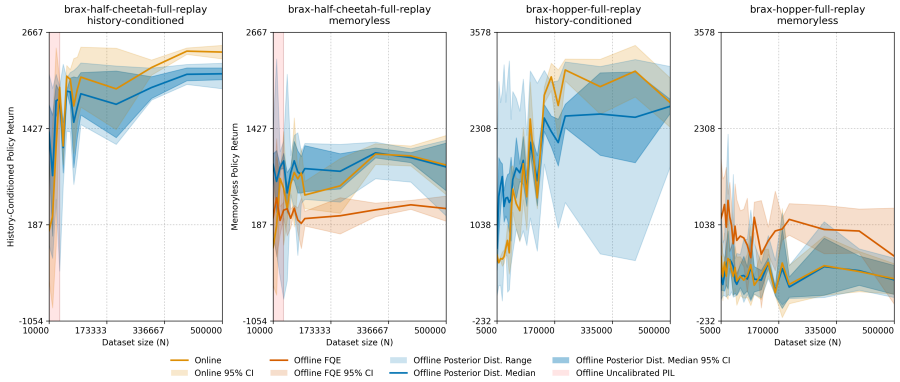


Figure 2: Evaluation of SOReL in zero-shot setting over 3 seeds. Policy hyperparameters are tuned **entirely offline** using the predictive value estimate. History is set to zero (memoryless) for comparison with FQE.

the value in terms of its absolute value. Instead, we treat the predictive value in Eq. (7) as a *value metric* that is positively correlated with true expected returns, and use this to tune ORL parameters ϕ_{III}. Our empirical evaluations in Section 7.2 support this hypothesis. We note that in model-free methods, the dynamics model and an approximate inference method are not used in policy learning, only to choose hyperparameters.

7 Experiments

The goal of our experiments is to confirm our methods provide a solution to **Issues I & II**: first, to evaluate SOReL's ability to approximate online performance entirely offline; and second, to evaluate TOReL's ability to tune hyperparameters of general ORL algorithms entirely offline.

7.1 Zero-Shot Fully Offline RL with Value Estimation

Our first experiment evaluates SOReL in the strict zero-shot setting: policy learning, hyperparameter tuning, and value estimation are performed using only offline data, with no interaction with the true environment prior to deployment. We assess whether the resulting offline value estimates reliably predict realised online performance. We test SOReL in five environments — two Gymnax (Lange, 2022) and three Brax (Freeman et al., 2021) — using datasets we collect to ensure broad behavioural diversity and state-action coverage (see Section F.1). To emulate practical usage, we progressively increase dataset size and estimate value entirely offline. We deploy each policy in the true environment to validate the estimated value, but in practice the policy would only be deployed once the estimated value is high enough and uncertainty small enough.

Results (Fig. 2 and Fig. 7) show that the posterior predictive median closely tracks realised online returns, which consistently lie within the predictive range induced by sampling dynamics from the posterior. Regions shaded in red (e.g., small datasets in Half-Cheetah) correspond to miscalibrated PIL, signalling unreliable estimates. Once the PIL is calibrated and predictive variance sufficiently reduced, deployment can be undertaken with confidence. The diagnostics are actionable. In Half-Cheetah, small datasets yield wide predictive ranges and poorly calibrated PIL, correctly warning against deployment; with more data, calibration improves and the median aligns tightly with online returns. In Hopper, although the PIL remains calibrated, predictive variance stays large, signalling persistent uncertainty. Crucially, this uncertainty is visible *offline*, enabling informed deployment decisions without online interaction. All SOReL hyperparameters are tuned entirely offline. As shown in Fig. 6 in Section E.1, predictive-value-based tuning substantially improves performance, with both Half-Cheetah ($r = 0.93, p = 0.00$) and Hopper ($r = 0.70, p = 0.00$) having statistically significant ($p < 0.05$), strong positive ($r > 0.5$) Pearson correlation between offline predictive value and realised online return. Moreover, SOReL learns policies that outperform the behaviour policies used to generate the datasets. For example, on a subset of the Half-Cheetah dataset, SOReL achieves nearly $1.5\times$ the return of the best data-collecting policy (Fig. 11), demonstrating its ability to improve beyond suboptimal data.

Benchmarking Value Estimation We benchmark SOReL's ability to approximate online performance with fitted Q-evaluation (FQE) as used in Paine et al. (2020), which is the closest existing algorithm to SOReL. We cannot benchmark against OPE metrics based on importance sampling

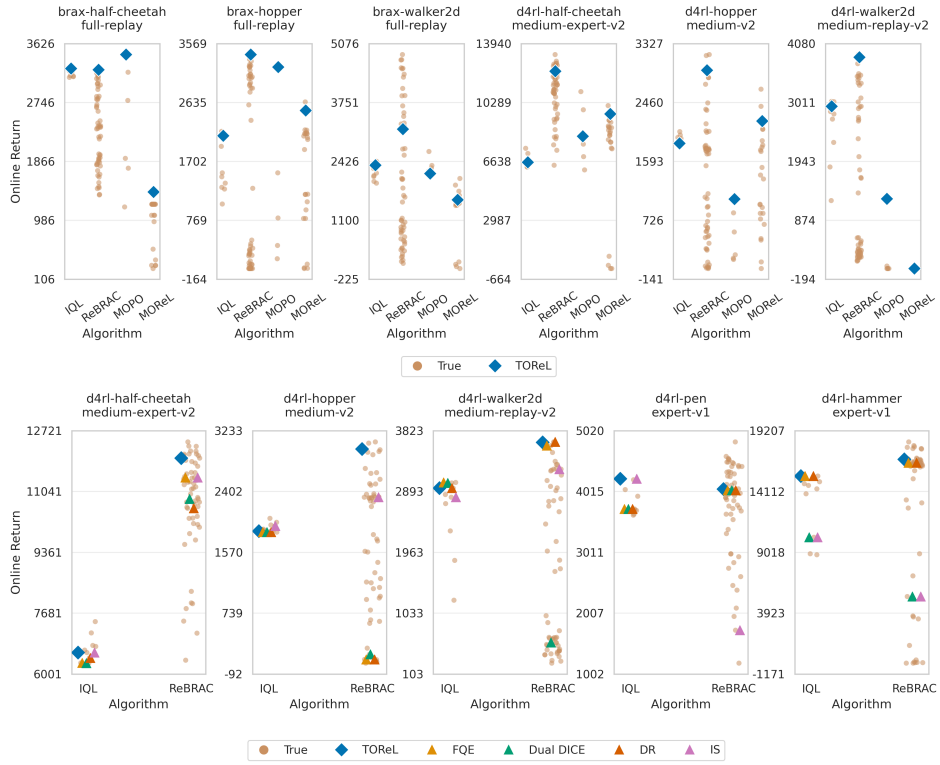


Figure 3: True online returns achieved by different hyperparameter combinations for different ORL algorithms and datasets. **Top**: hyperparameters selected by **TOReL** for IQL, ReBRAC, MOPO and MOReL. **Bottom:** hyperparameters selected by **TOReL** and 4 OPE metrics (**FQE**, **Dual DICE**, **DR** and **IS**) for IQL and ReBRAC. Error bars omitted due to the prohibitive cost of repeated runs (∼2000 experiments).

(Kostrikov & Nachum, 2020) or double robustness (Thomas & Brunskill, 2016), which require datasets containing trajectories rather than individual datapoints. As SOReL is a Bayesian approach, policies condition on histories to represent belief states. In contrast, FQE is designed for state-conditioned policies. For fair comparison, we reset the hidden states of all trajectories on each step to zero to project the history-conditioned policy to a state-conditioned policy. We compare SOReL's predictive value for the history-zeroed policy (as defined in Eq. (7)) with FQE's value estimate by evaluating both against the ground-truth returns of this policy.

Results (Fig. 2 in Section E.1) show that SOReL's predictive median tracks the realised online return more closely than FQE, with an average absolute error of 120 versus 470 for Half-Cheetah (over the points where the PIL is well-calibrated) and 70 versus 550 for Hopper. While FQE can approximate value in some regimes, it frequently exhibits overestimation or unstable behaviour for small datasets. In contrast, SOReL's posterior predictive range consistently contains the true return once the PIL is calibrated, and contracts as dataset size increases, reflecting improved data sufficiency. Moreover, when SOReL's predictive estimate deviates from the realised return, it tends to do so conservatively, underestimating rather than overestimating performance in low-data regimes. This pessimistic bias is desirable in safety-critical settings, where optimistic errors may lead to premature deployment. Unlike FQE, SOReL provides a calibrated uncertainty quantification alongside its value estimate; note that standard deviations are across policies, rather than for a single policy, like SOReL's return range is. The posterior predictive variance meaningfully reflects epistemic uncertainty: wide ranges for small datasets signal unreliable estimates, while contraction with additional data indicates increasing confidence. Finally, SOReL's PIL approach affords a principled diagnostic for detecting model misspecification, support mismatch, or insufficient data, while FQE does not provide an analogue of the PIL that links uncertainty calibration to regret or guides hyperparameter tuning.

These results highlight three key differences. Firstly, SOReL's predictive median approach yields more reliable value estimates in the zero-shot setting, with conservative behaviour when uncertainty is high. Secondly, SOReL provides actionable uncertainty diagnostics that allow practitioners to determine *offline* whether a value estimate is trustworthy. Thirdly, we remark that Paine et al. (2020)'s

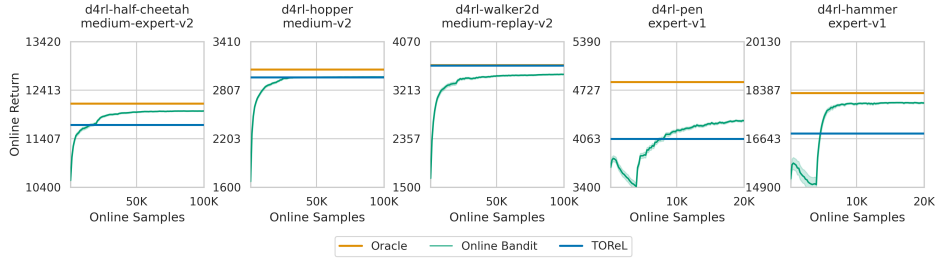


Figure 4: **TOReL**'s performance compared to an **online** bandit-based hyperparameter tuning algorithm. Shaded region indicates 95% bootstrap CI.

approach FQE still requires online hyperparameter tuning for its own hyperparameter set, whereas SOReL's approach is fully offline.

7.2 ORL Offline Hyperparameter Tuning

We carry out three experiments to test TOReL's ability as a fully offline hyperparameter tuning method for general ORL. Firstly, we apply TOReL to four existing ORL algorithms; two model-free, IQL (Kostrikov et al., 2021) and ReBRAC (Tarasov et al., 2023), and two model-based, MOPO (Yu et al., 2020) and MOReL Kidambi et al. (2020). We assess whether TOReL can identify the hyperparameter configuration yielding the highest true online return (Section 7.1). Second, we compare TOReL against OPE-based hyperparameter selection methods following Paine et al. (2020). Using five D4RL datasets (which contain full trajectories required by OPE methods), we evaluate IQL and ReBRAC under hyperparameters selected by TOReL and by FQE, Dual DICE, DR, and IS. As OPE algorithms have hyperparameters that are tuned online, we use the values given in Fu et al. (2021). We learn policies for the hyperparameter combinations suggested by Jackson et al. (2025). Results are shown in Section 7.1. Finally, we compare TOReL against Jackson et al. (2025)'s state-of-the-art online UCB-based hyperparameter tuning algorithm, measuring the online sample complexity required to replicate the performance obtained by TOReL's fully offline tuning (Fig. 4). Section 7.1 and Section 7.1 shows TOReL identifes hyperparameters with the high online returns across both different algorithms and datasets. Across all evaluated datasets, TOReL matches or exceeds the best online return achieved by hyperparameters selected via FQE, Dual DICE, DR, and IS. Moreover, TOReL's offline predictive metric exhibits strong positive correlation with true performance; in particular, it achieves statistically significant strong positive correlation on 18 of 28 hyperparameter tuning tasks (see Table 1 for Pearson correlations), and on 6 of 10 tasks where we benchmark against standard OPE metrics (compared to 4 for FQE, 0 for Dual DICE, and 1 each for DR and IS). Learning the ORL policy with ReBRAC and tuning hyperparameters with TOReL is a particularly strong combination. Finally, Fig. 4 highlights the sample-efficiency advantage of TOReL's fully offline tuning. From the spread of returns for different ReBRAC hyperparameter combinations, Fig. 4 demonstrates that online bandit-based tuning requires significant interaction to reliably identify high-performing settings. An additional 16000, 36000, 7700 and 5000 online samples are required to match TOReL's performance on Half-Cheetah, Hopper, Pen and Hammer, respectively. For Walker2d, TOReL's performance remains unmatched within the evaluated interaction budget.

8 Conclusion

High online sample complexity and lack of performance guarantees of existing methods present a major barrier to the widespread adoption of offline RL. In this paper, we introduce SOReL and TOReL, two theoretically grounded approaches to tackle these core issues. SOReL is a model-based Bayesian method that exploits predictive uncertainty to approximate online value entirely offline. Hyperparameters of the dynamics model and posterior are tuned by minimising and calibrating the PIL, while hyperparameters of the BAMDP solver are selected using predictive performance from posterior-sampled rollouts. TOReL extends this fully offline tuning procedure to general offline RL algorithms. Empirically, SOReL provides reliable offline performance estimates, and TOReL enables fully offline hyperparameter selection, substantially reducing online tuning requirements without sacrificing performance.

Broader impact: This paper presents work whose goal is to improve the safety and efficacy of offline RL. Our work therefore takes a step towards the development of safe offline RL methods with accurate value estimate that act predictably once deployed.

References

Alekh Agarwal, Nan Jiang, Sham M. Kakade, and Wen Sun. *Reinforcement Learning: Theory and Algorithms*. 2021. URL https://rltheorybook.github.io/.

J. Aitchison. Goodness of prediction fit. *Biometrika*, 62(3):547–554, 1975. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2335509.

Ahmed M. Alaa and Mihaela van der Schaar. Bayesian nonparametric causal inference: Information rates and learning algorithms. *IEEE Journal of Selected Topics in Signal Processing*, 12(5): 1031–1046, 2018. DOI: 10.1109/JSTSP.2018.2848230.

Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified q-ensemble. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 7436–7447. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/3d3d286a8d153a4a58156d0e02d8570c-Paper.pdf.

Mihaela Aslan. Asymptotically minimax bayes predictive densities. *The Annals of Statistics*, 34(6): 2921–2938, 2006. ISSN 00905364. URL http://www.jstor.org/stable/25463538.

Philip J Ball, Cong Lu, Jack Parker-Holder, and Stephen Roberts. Augmented world models facilitate zero-shot dynamics generalization from a single offline environment. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 619–629. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ball21a.html.

Andrew R Barron. Information-theoretic characterization of bayes performance and the choice of priors in parametric and nonparametric problems. In *Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting June 6-10, 1998*. Oxford University Press, 08 1999. ISBN 9780198504856. DOI: 10.1093/oso/9780198504856.003.0002. URL https://doi.org/10.1093/oso/9780198504856.003.0002.

A.R. Barron. *The Exponential Convergence of Posterior Probabilities with Implications for Bayes Estimators of Density Functions*. Department of Statistics, University of Illinois, 1988. URL https://books.google.co.uk/books?id=8raEnQAACAAJ.

R.F. Bass. *Real Analysis for Graduate Students*, chapter 21. Createspace Ind Pub, 2013. ISBN 9781481869140. URL https://books.google.co.uk/books?id=s6mVlgEACAAJ.

Jacob Beck, Risto Vuorio, Evan Zheran Liu, Zheng Xiong, Luisa Zintgraf, Chelsea Finn, and Shimon Whiteson. A survey of meta-reinforcement learning, 2024. URL https://arxiv.org/abs/2301.08028.

Richard Bellman. A problem in the sequential design of experiments. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 16(3/4):221–229, 1956. ISSN 00364452. URL http://www.jstor.org/stable/25048278.

Richard Bellman. Dynamic programming and stochastic control processes. *Information and Control*, 1(3):228–239, 1958. ISSN 0019-9958. DOI: https://doi.org/10.1016/S0019-9958(58)80003-0. URL https://www.sciencedirect.com/science/article/pii/S0019995858800030.

Blair Bilodeau, Dylan J. Foster, and Daniel M. Roy. Minimax rates for conditional density estimation via empirical entropy. *The Annals of Statistics*, 2021. URL https://api.semanticscholar.org/CorpusID:237592759.

J. L. Bretagnolle and Catherine Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 47:119–137, 1978. URL https://api.semanticscholar.org/CorpusID:122597694.

Jiayu Chen, Wentse Chen, and Jeff Schneider. Bayes adaptive monte carlo tree search for offline model-based reinforcement learning, 2024. URL https://arxiv.org/abs/2410.11234.

K Chua, R Calandra, R McAllister, and S Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. 2018.

Kamil Ciosek, Vincent Fortuin, Ryota Tomioka, Katja Hofmann, and Richard Turner. Conservative uncertainty estimation by fitting prior networks. In *Eighth International Conference on Learning Representations*, 04 2020.

B.S. Clarke and A.R. Barron. Information-theoretic asymptotics of bayes methods. *IEEE transactions on information theory*, 36(3):453–471, 1990. ISSN 0018-9448.

J. L. Doob. Application of the theory of martingales. In *Le calcul des probabilités et ses applications [The calculus of probabilities and its applications]*, number 13 in CNRS International Colloquia, pp. 23–27. Centre National de la Recherche Scientifique, Paris, 1949. (Lyon, France, 28 June–3 July 1948). MR:33460. Zbl:0041.45101.

Alvin Drake. Observation of a markov process through a noisy channel. *PhD Thesis*, 1962.

Yan Duan, John Schulman, Xi Chen, Peter Bartlett, Ilya Sutskever, and Peter Abbeel. Fast reinforcement learning via slow reinforcement learning. 1987.

Michael O'Gordon Duff. *Optimal Learning: Computational Procedures for Bayes-Adaptive Markov Decision Processes*. PhD thesis, 2002. AAI3039353.

Mattie Fellows, Brandon Kaplowitz, Christian Schroeder de Witt, and Shimon Whiteson. Bayesian exploration networks. In *ICML*, 2024.

C. Daniel Freeman, Erik Frey, Anton Raichuk, Sertan Girgin, Igor Mordatch, and Olivier Bachem. Brax – a differentiable physics engine for large scale rigid body simulation, 2021. URL https://arxiv.org/abs/2106.13281.

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

Justin Fu, Mohammad Norouzi, Ofir Nachum, George Tucker, Ziyu Wang, Alexander Novikov, Mengjiao Yang, Michael R. Zhang, Yutian Chen, Aviral Kumar, Cosmin Paduraru, Sergey Levine, and Tom Le Paine. Benchmarks for deep off-policy evaluation, 2021. URL https://arxiv.org/abs/2103.16596.

Scott Fujimoto and Shixiang (Shane) Gu. A minimalist approach to offline reinforcement learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 20132–20145. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/a8166da05c5a094f7dc03724b41886e5-Paper.pdf.

Arthur Guez, David Silver, and Peter Dayan. Efficient bayes-adaptive reinforcement learning using sample-based search. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/35051070e572e47d2c26c241ab88307f-Paper.pdf.

Arthur Guez, David Silver, and Peter Dayan. Scalable and efficient bayes-adaptive reinforcement learning based on monte-carlo tree search. *Journal of Artificial Intelligence Research*, 48:841–883, 10 2013. DOI: 10.1613/jair.4117.

Arthur Guez, Nicolas Heess, David Silver, and Peter Dayan. Bayes-adaptive simulation-based search with value function approximation. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/74d863ca4a12ccca50a754a3b277dbf7-Paper.pdf.

J. A. Hartigan. The maximum likelihood prior. *The Annals of statistics*, 26(6):2083–2103, 1998. ISSN 0090-5364.

Matthew Thomas Jackson, Uljad Berdica, Jarek Liesen, Shimon Whiteson, and Jakob Nicolaus Foerster. A clean slate for offline reinforcement learning. *arXiv preprint arXiv:2504.11453*, 2025.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artif. Intell.*, 101(1–2):99–134, may 1998. ISSN 0004-3702.

Robert E Kass, Luke Thierney, and Joeseph B Kadane. The validity of posterior expansions based on laplace's method. *Bayesian and Likelihood Methods in Statistics and Economics*, pp. 473–488, 1990. URL https://www.stat.cmu.edu/~kass/papers/validity.pdf.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21810–21823. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f7efa4f864ae9b88d43527f4b14f750f-Paper.pdf.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Fumiyasu Komaki. On asymptotic properties of predictive distributions. *Biometrika*, 83(2):299–313, 06 1996. ISSN 0006-3444. DOI: 10.1093/biomet/83.2.299. URL https://doi.org/10.1093/biomet/83.2.299.

Ilya Kostrikov and Ofir Nachum. Statistical bootstrapping for uncertainty estimation in off-policy evaluation, 2020. URL https://arxiv.org/abs/2007.13609.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *CoRR*, abs/2110.06169, 2021. URL https://arxiv.org/abs/2110.06169.

Danie G. Krige. A statistical approach to some basic mine valuation problems on the witwatersand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, 52:119–139, 1951.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/0d2b2061826a5df3221116a5085a6052-Paper.pdf.

Robert Tjarko Lange. gymnax: A JAX-based reinforcement learning environment library, 2022. URL http://github.com/RobertTLange/gymnax.

Sascha Lange, Thomas Gabel, and Martin Riedmiller. *Batch Reinforcement Learning*, pp. 45–73. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-27645-3. DOI: 10.1007/978-3-642-27645-3_2. URL https://doi.org/10.1007/978-3-642-27645-3_2.

Hoang M. Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints, 2019. URL https://arxiv.org/abs/1903.08738.

Lucien Le Cam. On some asymptotic properties of maximum likelihood estimates and related bayes' estimates. volume 1, pp. 277–300, 1953.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems, 2020. URL https://arxiv.org/abs/2005.01643.

Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *ArXiv*, abs/2204.05275, 2022. URL https://arxiv.org/pdf/2204.05275.

D. V. Lindley. Approximate bayesian methods. *Trabajos de Estadistica Y de Investigacion Operativa*, 31(1):223–245, 1980.

Chris Lu, Jakub Kuba, Alistair Letcher, Luke Metz, Christian Schroeder de Witt, and Jakob Foerster. Discovered policy optimisation. *Advances in Neural Information Processing Systems*, 35:16455–16468, 2022a.

Cong Lu, Philip Ball, Jack Parker-Holder, Michael Osborne, and Stephen J. Roberts. Revisiting design choices in offline model based reinforcement learning. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=zz9hXVhf40.

J. J. Martin. *Bayesian decision problems and Markov chains [by] J. J. Martin*. Wiley New York, 1967.

Kevin Murphy. Reinforcement learning: An overview, 2024. URL https://arxiv.org/abs/2412.05265.

O Nachum, Y Chow, B Dai, and L Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. 2019.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2701–2710. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/osband17a.html.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 8617–8629. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/8080-randomized-prior-functions-for-deep-reinforcement-learning.pdf.

Tom Le Paine, Cosmin Paduraru, Andrea Michi, Çaglar Gülçehre, Konrad Zolna, Alexander Novikov, Ziyu Wang, and Nando de Freitas. Hyperparameter selection for offline reinforcement learning. *CoRR*, abs/2007.09055, 2020. URL https://arxiv.org/abs/2007.09055.

K. B. Petersen and M. S. Pedersen. The matrix cookbook, nov 2012. URL http://localhost/pubdb/p.php?3274. Version 20121115.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, 1994. ISBN 0471619779.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006. URL https://gaussianprocess.org/gpml/.

Gareth O. Roberts and Jeffrey S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1(none):20 – 71, 2004. DOI: 10.1214/154957804100000024. URL https://doi.org/10.1214/154957804100000024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

Anya Sims, Cong Lu, Jakob Nicolaus Foerster, and Yee Whye Teh. The edge-of-reach problem in offline model-based reinforcement learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=3dn1hINA6o.

Richard D. Smallwood and Edward J. Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973. ISSN 0030364X, 15265463. URL http://www.jstor.org/stable/168926.

Matthew Smith, Lucas Maystre, Zhenwen Dai, and Kamil Ciosek. A strong baseline for batch imitation learning, 2023. URL https://arxiv.org/abs/2302.02788.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Scholkopf, and Gert R. G. Lanckriet. On integral probability metrics, ϕ-divergences and binary classification. *arXiv: Information Theory*, 2009. URL https://api.semanticscholar.org/CorpusID:14114329.

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

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

Denis Tarasov, Vladislav Kurenkov, Alexander Nikulin, and Sergey Kolesnikov. Revisiting the minimalist approach to offline reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=vqGWslLeEw.

Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning, 2016. URL https://arxiv.org/abs/1604.00923.

Luke Tierney and Joseph B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986. ISSN 0162-1459.

Luke Tierney, Robert E. Kass, and Joseph B. Kadane. Fully exponential laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, 84(407):710–716, 1989. ISSN 0162-1459.

Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning, 2022. URL https://arxiv.org/abs/2212.06355.

A. W. van der Vaart. *Bayes Procedures*, pp. 138–152. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. DOI: 10.1017/CBO9780511802256.011.

Aad van der Vaart and Harry van Zanten. Information rates of nonparametric gaussian process methods. *J. Mach. Learn. Res.*, 12(null):2095–2119, July 2011. ISSN 1532-4435.

Han Wang, Archit Sakhadeo, Adam M White, James M Bell, Vincent Liu, Xutong Zhao, Puer Liu, Tadashi Kozuno, Alona Fyshe, and Martha White. No more pesky hyperparameters: Offline hyperparameter tuning for RL. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=AiOUi3440V.

Ruosong Wang, Dean P. Foster, and Sham M. Kakade. What are the statistical limits of offline rl with linear function approximation?, 2020. URL https://arxiv.org/abs/2010.11895.

Norbert Wiener. Differential-space. *Journal of Mathematics and Physics*, 2(1-4):131–174, 1923. DOI: https://doi.org/10.1002/sapm192321131. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sapm192321131.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564 – 1599, 1999. DOI: 10.1214/aos/1017939142. URL https://doi.org/10.1214/aos/1017939142.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14129–14142. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/a322852ce0df73e204b7e67cbbef0d0a-Paper.pdf.

Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. Combo: Conservative offline model-based policy optimization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 28954–28967. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f29a179746902e331572c483c45e5086-Paper.pdf.

Baohe Zhang, Raghu Rajan, Luis Pineda, Nathan Lambert, André Biedenkapp, Kurtland Chua, Frank Hutter, and Roberto Calandra. On the importance of hyperparameter optimization for model-based reinforcement learning. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 4015–4023. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/zhang21n.html.

Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/pdf?id=Hkl9JlBYvr.

Åström, Karl Johan. Optimal Control of Markov Processes with Incomplete State Information I. 10:174–205, 1965. ISSN 0022-247X. DOI: {10.1016/0022-247X(65)90154-X}. URL https://lup.lub.lu.se/search/files/5323668/8867085.pdf.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Primer on Bayesian RL

A Bayesian epistemology characterises the agent's uncertainty in the MDP through distributions over any unknown variable (Martin, 1967; Duff, 2002). In our learning problem, a Bayesian first specifies a model $P_{R,S}(s_t, a_t)$ over the unknown state transition and reward distribution, representing a hypothesis space of possible environment dynamics. We focus on a parametric model: $p(r_t, s_{t+1}|s_t, a_t, \theta)$ with each $\theta \in \Theta \subseteq \mathbb{R}^d$ representing a hypothesis about the MDP $\mathcal{M}^\star$, however our results can easily be generalised to non-parametric methods like Gaussian process regression (Rasmussen & Williams, 2006; Wiener, 1923; Krige, 1951). A prior distribution over the parameter space P_Θ is specified, which represents the initial *a priori* belief in the true value of $P^\star_{R,S}(s, a)$ before the agent has observed any transitions. Priors are a powerful aspect of Bayesian RL, allowing practitioners to provide the agent with any information about the MDP and transfer knowledge between agents and domains. Given a history h_t, the prior is updated to a posterior $P_\Theta(h_t)$, representing the agent's beliefs in the MDP's dynamics once h_t has been observed. For each history, the posterior is used to *marginalise* across all hypotheses according to the agent's uncertainty, yielding the predictive state transition-reward distribution $P_{R,S}(h_t, a_t) = \mathbb{E}_{\theta\sim P_\Theta(h_t)}\left[P_{R,S}(s_t, a_t, \theta)\right]$ which characterise the epistemic and aleatoric uncertainty in $P^\star_{R,S}(s_t, a_t)$. Given $P_{R,S}(h_t, a_t)$, we reason over counterfactual future trajectories using the predictive distribution over trajectories P^π_t and define the BRL objective as:

$$J^\pi_{\text{Bayes}}(P_\Theta) \coloneqq \mathbb{E}_{h_\infty\sim P^\pi_\infty}\left[\sum_{i=0}^{\infty}\gamma^i r_i\right] = \mathbb{E}_{\theta\sim P_\Theta}\left[\mathbb{E}_{h_\infty\sim P^\pi_\infty(\theta)}\left[\sum_{i=0}^{\infty}\gamma^i r_i\right]\right].$$

Let $\Pi_\mathcal{H}$ denote the space of all history-conditioned policies. A corresponding optimal policy is known as a Bayes-optimal policy, which we denote as $\pi^\star_{\text{Bayes}}(\cdot) \in \Pi^\star_{\text{Bayes}}(P_\Theta) \coloneqq \arg\max_{\pi\in\Pi_\mathcal{H}} J^\pi_{\text{Bayes}}(P_\Theta)$. Unlike in frequentist RL, Bayesian variables depend on histories obtained through posterior marginalisation; hence the posterior is often known as the *belief state*, which augments each ground state s_t like in a partially observable Markov decision process (Drake, 1962; Åström, Karl Johan, 1965; Smallwood & Sondik, 1973; Kaelbling et al., 1998). Analogously to the state-transition distribution in frequentist RL, we define a *belief transition* distribution $P_\mathcal{H}(h_t, a_t)$ using the predictive state transition-reward distribution, which yields *Bayes-adaptive MDP* (BAMDP) (Duff, 2002): $\mathcal{M}_{\text{Bayes}}(P_\Theta) \coloneqq \langle\mathcal{H}, \mathcal{A}, P_0, P_\mathcal{H}(h, a), \gamma\rangle$. The BAMDP is solved using planning methods to obtain a Bayes-optimal policy, which naturally balances exploration with exploitation: after every timestep, the agent's uncertainty is characterised via the posterior conditioned on the history h_t, which includes all future trajectories to marginalise over. Via the belief transition, the BRL objective accounts for how the posterior evolves on every timestep, and hence any Bayes-optimal policy $\pi^\star_{\text{Bayes}}$ is optimal not only according to the epistemic uncertainty of a fixed belief but accounts for how the epistemic uncertainty evolves at every future timestep, decaying according to the discount factor.

Let

$$J^\pi(\theta) \coloneqq \mathbb{E}_{h_\infty\sim P^\pi_\infty(\theta)}\left[\sum_{i=0}^{\infty}\gamma^i r_i\right],$$

denote the expected returns under policy π for an MDP parametrised by θ and $J^\star(\theta) \coloneqq \sup_\pi J^\pi(\theta)$ denote the expected returns for under an optimal policy or an MDP parametrised by θ. The Bayes regret of a policy π is defined as:

$$\begin{aligned}\text{Regret}^\pi_{\text{Bayes}}(P_\Theta) &\coloneqq \mathbb{E}_{\theta\sim P_\Theta}\left[J^\star(\theta) - J^\pi(\theta)\right],\\ &= \mathbb{E}_{\theta\sim P_\Theta}\left[J^\star(\theta)\right] - J^\pi_{\text{Bayes}}(P_\Theta).\end{aligned} \tag{8}$$

which characterises how far a policy π is from the true optimal policy taking an average over all possible MDPs using the prior P_Θ. It is clear from Eq. (8) that any policy that minimises Bayes regret maximises the Bayesian RL objective $J^\pi_{\text{Bayes}}(P_\Theta)$.

B Derivation of Negative Log Likelihood Loss Function with Approximate Inference

Assume a dataset of N input-output pairs:

$$\mathcal{D}_N \coloneqq \{(x_0, y_0), (x_1, y_1), \dots (x_{N-1}, y_{N-1})\},$$

and a multivariate Gaussian regression model:

$$p(y|x,\theta) = \frac{1}{(2\pi)^{\frac{D}{2}}} |\Sigma_\theta(x)|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(y-\mu_\theta(x))\Sigma_\theta^{-1}(y-\mu_\theta(x))^T\right),$$

where D is the number of dimensions. Here our model $\text{NN}_\theta : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ is a neural network parametrised by $\theta \in \Theta$ that outputs a Gaussian distribution $\mathcal{N}(\mu_\theta, \Sigma_\theta)$ over $\mathcal{Y}$. Assuming independent dimensions, such that the covariance matrix is diagonal:

$$p(y|x,\theta) = \prod_{d=0}^{D-1} \frac{1}{\sqrt{2\pi\sigma^2_{\theta_d}(x)}} \exp\left(-\frac{1}{2\sigma^2_{\theta_d}(x)}(y_d - \mu_{\theta_d}(x))^2\right).$$

We can then fit our model by minimising the negative log likelihood loss:

$$\begin{aligned}
&\mathcal{L}(\text{NLL}(\theta)) \\
&\quad \coloneqq -\log p(\mathcal{D}_N|\theta), \\
&\quad = \sum_{i=0}^{N-1}\left(\frac{D}{2}\log(2\pi) + \frac{1}{2}\sum_{d=0}^{D-1}\left(\log\sigma^2_{\theta_d}(x_i) + \frac{(y_{i_d} - \mu_{\theta_d}(x_i))^2}{\sigma^2_{\theta_d}(x_i)}\right)\right), \\
&\quad = \left[\sum_{i=0}^{N-1}\frac{1}{N}\left(\frac{D}{2}\log(2\pi) + \frac{1}{2}\sum_{d=0}^{D-1}\left(\log\sigma^2_{\theta_d}(x_i) + \frac{(y_{i_d} - \mu_{\theta_d}(x_i))^2}{\sigma^2_{\theta_d}(x_i)}\right)\right)\right], \\
&\quad = \mathbb{E}_{i\sim\mathcal{U}_N}\left[\frac{D}{2}\log(2\pi) + \frac{1}{2}\sum_{d=0}^{D-1}\left(\log\sigma^2_{\theta_d}(x_i) + \frac{(y_{i_d} - \mu_{\theta_d}(x_i))^2}{\sigma^2_{\theta_d}(x_i)}\right)\right], \\
&\quad \overset{c}{=} \mathbb{E}_{i\sim\mathcal{U}_N}\left[\sum_{d=0}^{D-1}\left(\log\sigma^2_{\theta_d}(x_i) + \frac{(y_{i_d} - \mu_{\theta_d}(x_i))^2}{\sigma^2_{\theta_d}(x_i)}\right)\right],
\end{aligned}$$

where recall $\mathcal{U}_N$ is the uniform distribution over $\{0, 1, \dots N-1\}$. Our final line means equality up to a constant, as we can ignore the $\frac{D}{2}\log(2\pi)$ term for optimisation because it is independent of θ.

We use RP ensembles for our approximate posterior (Osband et al., 2018; Ciosek et al., 2020); here an ensemble of M separate model weights $\{\theta_0, \theta_i, \dots \theta_{M-1}\}$ are randomly initialised and are optimised in parallel, summing over the corresponding negative log likelihoods. When training, we optimise the log-variance rather than the variance for numerical stability and to ensure that the variance remains positive. This allows us to simultaneously optimise maximum and minimum log-variance parameters for each dimension across the ensemble, which we use to soft-clamp the log-variances output by individual models, preventing any individual model becoming overly confident or too uncertain in

one dimension. Our final loss function is then given by:

$$\mathcal{L}(\theta, \mathcal{D}_N) = \sum_{j=0}^{M-1} \left(\mathbb{E}_{i \sim \mathcal{U}_N} \left[\sum_{d=0}^{D-1} \left(\xi_{\theta_{d_j}}(x_i) + \frac{(y_{i_d} - \mu_{\theta_{d_j}}(x_i))^2}{\exp\left(\xi_{\theta_{d_j}}(x_i)\right)} \right) \right] \right) + c \cdot \sum_{d=0}^{D-1} \left(\xi_{\theta_{d_{\max}}} - \xi_{\theta_{d_{\min}}} \right),$$

where $\xi_{\theta_{d_j}} = \xi_{\theta_{d_{\min}}} + \left[1 + \exp\left(\xi_{\theta_{d_{\max}}} - \left[1 + \exp\left(\log \sigma^2_{\theta_{d_j}} - \xi_{\theta_{d_{\max}}}\right)\right] - \xi_{\theta_{d_{\min}}}\right)\right]$ and $\xi_{\theta_{d_{\min}}}$ and $\xi_{\theta_{d_{\max}}}$ are respectively the minimum and maximum log-variance parameters optimised across the ensemble, c is the log-variance difference coefficient used to control the clamping term, and M is the number of models in the ensemble. $\mathcal{L}(\theta, \mathcal{D}_N)$ can be minimised by using Monte Carlo minibatch gradient descent with a minibatch $\mathcal{M}_n$ of $n < N$ samples drawn uniformly from $\mathcal{D}_N$.

C Value Estimators

To estimate the value of the Bayes-optimal policy using only offline data, we sample dynamics models from the posterior (approximated via an ensemble), roll out $\pi^\star_{\text{Bayes}}$ in each sampled model, and compute the corresponding discounted returns. This yields predictive returns

$$\{R_\theta^{\pi^\star_{\text{Bayes}}}\}_{\theta \sim P_\Theta(\mathcal{D}_N)},$$

which approximate the posterior predictive distribution over the policy value. From these samples, we construct estimators of $J^{\pi^\star_{\text{Bayes}}}(\mathcal{M}^\star)$.

Specifically, we define the predictive mean:

$$\hat{J}^{\pi^\star_{\text{Bayes}}}_{\text{mean}}(\mathcal{M}^\star) = \mathbb{E}_{\theta \sim P_\Theta(\mathcal{D}_N)} \left[R_\theta^{\pi^\star_{\text{Bayes}}} \right],$$

the predictive median:

$$\hat{J}^{\pi^\star_{\text{Bayes}}}_{\text{median}}(\mathcal{M}^\star) = \mathbb{M}_{\theta \sim P_\Theta(\mathcal{D}_N)} \left[R_\theta^{\pi^\star_{\text{Bayes}}} \right],$$

the predictive maximum:

$$\hat{J}^{\pi^\star_{\text{Bayes}}}_{\max}(\mathcal{M}^\star) = \max_{\theta \sim P_\Theta(\mathcal{D}_N)} R_\theta^{\pi^\star_{\text{Bayes}}},$$

and the predictive minimum

$$\hat{J}^{\pi^\star_{\text{Bayes}}}_{\min}(\mathcal{M}^\star) = \min_{\theta \sim P_\Theta(\mathcal{D}_N)} R_\theta^{\pi^\star_{\text{Bayes}}}.$$

These estimators reflect different attitudes toward posterior uncertainty. The maximum is optimistic and can overestimate performance if even a single posterior sample is favourable, while the minimum is conservative and may underestimate performance when a subset of posterior samples are pessimistic. The predictive mean aggregates all samples but can be sensitive to extreme values. Empirically, we find that the predictive median provides the most stable trade-off: it is robust to outliers, avoids undue optimism driven by a small number of favourable models, and is less conservative than the minimum. In practice, $\hat{J}^{\pi^\star_{\text{Bayes}}}_{\text{median}}(\mathcal{M}^\star)$ tracks the true online value closely across domains, making it a reliable fully offline estimator.

D Proofs

D.1 Primer on Total Variational Distance

We measure distance between two probability distributions P_X and Q_X using the total variational (TV) distance, defined as:

$$\mathrm{TV}(P_X \| Q_X) \coloneqq \sup_E |P_X(E) - Q_X(E)| .$$

The TV distance takes the supremum over all events E to find the event that gives rise to the maximum difference in probability between two distributions. A key property of the TV distance is: $0 \leq \mathrm{TV}(P_X \| Q_X) \leq 1$. If $\mathrm{TV}(P_X \| Q_X) = 0$, then $P_X = Q_X$ as there is no event that both distributions don't assign the same probability mass to. If $\mathrm{TV}(P_X \| Q_X) = 1$, then the distributions assign completely different mass to at least one event. The TV distance can be related to the Kullback-Leibler (KL) divergence using the Bretagnolle-Huber (Bretagnolle & Huber, 1978) inequality: $\mathrm{TV}(P_X \| Q_X) \leq \sqrt{1 - \exp\left(-\mathrm{KL}(P_X \| Q_X)\right)} \leq 1$, which preserves the property that $0 \leq \mathrm{TV}(P_X \| Q_X) \leq 1$. The TV distance can be shown (Sriperumbudur et al., 2009) to be equivalent to the integral probability metric under the ∞-norm, which we will make use of in our theorems:

$$\mathrm{TV}(P_X \| Q_X) = \frac{1}{2} \sup_{f \in \mathcal{F}: \mathcal{X} \to [-1,1]} \left| \mathbb{E}_{x \sim P_X}\left[f(x)\right] - \mathbb{E}_{x \sim Q_X}\left[f(x)\right] \right| , \tag{9}$$

In this form, the supremum is taken over the space of all functions that are bounded by unity, that is $\|f\|_\infty = 1$.

D.2 Proof of Theorem 1

Let the predictive distribution over history h_t using the posterior $P_\Theta(\mathcal{D}_N)$ be $P_{t,\pi}(\mathcal{D}_N)$, which has density:

$$p_\pi(h_t, \mathcal{D}_N) \coloneqq p_0(s_0) \prod_{i=0}^{t-1} \pi(a_i|h_i) p(r_i|h_i, a_i, \mathcal{D}_N) p(s_{i+1}|h_i, a_i, \mathcal{D}_N).$$

According to the Bernstein-von Mises theorem (Doob, 1949; Le Cam, 1953; Vaart, 1998), as the posterior becomes more informative it concentrates around a smaller (and more tractable) subset of the hypothesis space. Not only does this ease the computational burden of solving the BRL objective, but in the limit $N \to \infty$, the Bayesian RL objective using the true posterior will approach the true expected discounted return for the MDP: $J^\pi(P_\Theta(\mathcal{D}_N)) \xrightarrow[N \to \infty]{} J^\pi(\mathcal{M}^\star)$. In this limit, any Bayes-optimal policy will be an optimal policy for the true MDP, achieving the highest expected returns once deployed. Our first lemma bounds the difference between the expected true rewards after time i and the predictive expected rewards in terms of total variational distance:

Lemma 1. *For bounded reward functions:*

$$\left| \mathbb{E}_{h_{i+1} \sim P^\star_{i+1,\pi}}\left[r_i\right] - \mathbb{E}_{h_{i+1} \sim P^\pi_{i+1}(\mathcal{D}_N)}\left[r_i\right] \right| \leq (r_{\max} - r_{\min}) \cdot \mathrm{TV}\left(P^\star_{i+1,\pi} \| P^\pi_{i+1}(\mathcal{D}_N)\right) . \tag{10}$$

Proof. We start by subtracting and adding $\frac{r_{\max}+r_{\min}}{2}$ to the left hand side of Ineq. 10:

$$\begin{aligned}
&|\mathbb{E}_{h_{i+1} \sim P^\star_{i+1,\pi}}\left[r_i\right] - \mathbb{E}_{h_{i+1} \sim P^\pi_{i+1}(\mathcal{D}_N)}\left[r_i\right]| \\
&= \left| \mathbb{E}_{h_{i+1} \sim P^\star_{i+1,\pi}}\left[r_i - \frac{r_{\max} + r_{\min}}{2} \right] - \mathbb{E}_{h_{i+1} \sim P^\pi_{i+1}(\mathcal{D}_N)}\left[r_i - \frac{r_{\max} + r_{\min}}{2} \right] \right| , \\
&= \left| \mathbb{E}_{h_{i+1} \sim P^\star_{i+1,\pi}}\left[\frac{2r_i - (r_{\max} + r_{\min})}{2} \right] - \mathbb{E}_{h_{i+1} \sim P^\pi_{i+1}(\mathcal{D}_N)}\left[\frac{2r_i - (r_{\max} + r_{\min})}{2} \right] \right| , \\
&= \frac{(r_{\max} - r_{\min})}{2} \cdot \left| \mathbb{E}_{h_{i+1} \sim P^\star_{i+1,\pi}}\left[\frac{2r_i - (r_{\max} + r_{\min})}{r_{\max} - r_{\min}} \right] - \mathbb{E}_{h_{i+1} \sim P^\pi_{i+1}(\mathcal{D}_N)}\left[\frac{2r_i - (r_{\max} + r_{\min})}{r_{\max} - r_{\min}} \right] \right| , \\
&= \frac{(r_{\max} - r_{\min})}{2} \cdot \left| \mathbb{E}_{h_{i+1} \sim P^\star_{i+1,\pi}}\left[r_{\mathrm{norm}}(h_{i+1})\right] - \mathbb{E}_{h_{i+1} \sim P^\pi_{i+1}(\mathcal{D}_N)}\left[r_{\mathrm{norm}}(h_{i+1})\right] \right| ,
\end{aligned} \tag{11}$$

where:

$$r_{\text{norm}}(h_{i+1}) \coloneqq \frac{2r_i - (r_{\max} + r_{\min})}{r_{\max} - r_{\min}}.$$

Now, as $r_{\text{norm}} : \mathcal{H}_{i+1} \to [-1, 1]$, we can bound Eq. (11) using the integral probability metric form of the TV distance (see Eq. (9)), yielding our desired result:

$$\begin{aligned}
&|\mathbb{E}_{h_{i+1}\sim P^{\star}_{i+1,\pi}}[r_i] - \mathbb{E}_{h_{i+1}\sim P^{\pi}_{i+1}(\mathcal{D}_N)}[r_i]| \\
&\leq \frac{r_{\max} - r_{\min}}{2} \cdot \sup_{f\in\mathcal{F}:\mathcal{H}_{i+1}\to[-1,1]} \left|\mathbb{E}_{h_{i+1}\sim P^{\star}_{i+1,\pi}}[f(h_{i+1})] - \mathbb{E}_{h_{i+1}\sim P^{\pi}_{i+1}(\mathcal{D}_N)}[f(h_{i+1})]\right|, \\
&= (r_{\max} - r_{\min}) \cdot \text{TV}\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\mathcal{D}_N)\right).
\end{aligned}$$

□

Using this result, we bound the true regret $\text{Regret}(\mathcal{M}^{\star}, \mathcal{D}_N)$ using the TV distance between the true $P^{\star}_{i,\pi}$ and predictive $P_{i,\pi}(\mathcal{D}_N)$ history distributions:

Lemma 2. *Let $\mathcal{R}_{\max} \coloneqq \frac{(r_{max} - r_{min})}{1-\gamma}$ denote the maximum possible regret for the MDP. For a prior $P_{\Theta}(\mathcal{D}_N)$, the true regret can be bounded as:*

$$\text{Regret}(\mathcal{M}^{\star}, \mathcal{D}_N) \leq 2\mathcal{R}_{\max} \cdot \sup_{\pi} \mathbb{E}_{i\sim\mathcal{G}(\gamma)}\left[TV\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\mathcal{D}_N)\right)\right]. \tag{12}$$

Proof. We start from the definition of the true regret:

$$\text{Regret}(\mathcal{M}^{\star}, \mathcal{D}_N) \coloneqq J^{\pi^{\star}}(\mathcal{M}^{\star}) - J^{\pi^{\star}_{\text{Bayes}}}(\mathcal{M}^{\star}, \mathcal{D}_N).$$

We now bound the difference between $J^{\pi^{\star}}(\mathcal{M}^{\star})$ and $J^{\pi^{\star}_{\text{Bayes}}}(\mathcal{M}^{\star}, \mathcal{D}_N)$ in terms of the difference between $J^{\pi}(\mathcal{M}^{\star})$ and $J^{\pi}_{\text{Bayes}}(\mathcal{D}_N)$:

$$\begin{aligned}
&\text{Regret}(\mathcal{M}^{\star}, \mathcal{D}_N) \\
&= J^{\pi^{\star}}(\mathcal{M}^{\star}) - J^{\pi^{\star}}_{\text{Bayes}}(P_{\Phi}(\mathcal{D}_N)) + J^{\pi^{\star}}_{\text{Bayes}}(P_{\Phi}(\mathcal{D}_N)) - J^{\pi^{\star}_{\text{Bayes}}}(\mathcal{M}^{\star}, \mathcal{D}_N), \\
&\leq J^{\pi^{\star}}(\mathcal{M}^{\star}) - J^{\pi^{\star}}_{\text{Bayes}}(P_{\Phi}(\mathcal{D}_N)) + J^{\pi^{\star}_{\text{Bayes}}}_{\text{Bayes}}(P_{\Phi}(\mathcal{D}_N)) - J^{\pi^{\star}_{\text{Bayes}}}(\mathcal{M}^{\star}, \mathcal{D}_N), \\
&\leq \sup_{\pi} \left|J^{\pi}(\mathcal{M}^{\star}) - J^{\pi}_{\text{Bayes}}(P_{\Theta}(\mathcal{D}_N))\right| + \sup_{\pi} \left|J^{\pi}_{\text{Bayes}}(P_{\Theta}(\mathcal{D}_N)) - J^{\pi}(\mathcal{M}^{\star})\right|, \\
&= 2\sup_{\pi} \left|J^{\pi}(\mathcal{M}^{\star}) - J^{\pi}_{\text{Bayes}}(P_{\Theta}(\mathcal{D}_N))\right|,
\end{aligned} \tag{13}$$

where the second line follows from $J^{\pi^{\star}}_{\text{Bayes}}(P_{\Phi}(\mathcal{D}_N)) \leq J^{\pi^{\star}_{\text{Bayes}}}_{\text{Bayes}}(P_{\Phi}(\mathcal{D}_N))$ by definition. Now our goal is to bound $\left|J^{\pi}(\mathcal{M}^{\star}) - J^{\pi}_{\text{Bayes}}(P_{\Theta}(\mathcal{D}_N))\right|$:

$$\begin{aligned}
\left|J^{\pi}(\mathcal{M}^{\star}) - J^{\pi}_{\text{Bayes}}(P_{\Theta}(\mathcal{D}_N))\right| &= \left|\mathbb{E}_{h_{\infty}\sim P^{\star}_{\infty,\pi}}\left[\sum_{i=0}^{\infty}\gamma^i r_i\right] - \mathbb{E}_{h_{\infty}\sim P^{\pi}_{\infty}(\mathcal{D}_N)}\left[\sum_{i=0}^{\infty}\gamma^i r_i\right]\right|, \\
&= \left|\sum_{i=0}^{\infty}\gamma^i \mathbb{E}_{h_{i+1}\sim P^{\star}_{i+1,\pi}}[r_i] - \sum_{i=0}^{\infty}\gamma^i \mathbb{E}_{h_{i+1}\sim P^{\pi}_{i+1}(\mathcal{D}_N)}[r_i]\right|, \\
&= \left|\sum_{i=0}^{\infty}\gamma^i \left(\mathbb{E}_{h_{i+1}\sim P^{\star}_{i+1,\pi}}[r_i] - \mathbb{E}_{h_{i+1}\sim P^{\pi}_{i+1}(\mathcal{D}_N)}[r_i]\right)\right|, \\
&\leq \sum_{i=0}^{\infty}\gamma^i \left|\mathbb{E}_{h_{i+1}\sim P^{\star}_{i+1,\pi}}[r_i] - \mathbb{E}_{h_{i+1}\sim P^{\pi}_{i+1}(\mathcal{D}_N)}[r_i]\right|.
\end{aligned}$$

Using Ineq. 10 from Lemma 1, we now bound each difference $\left|\mathbb{E}_{h_{i+1}\sim P^{\star}_{i+1,\pi}}[r_i] - \mathbb{E}_{h_{i+1}\sim P^{\pi}_{i+1}(\mathcal{D}_N)}[r_i]\right|$ in terms of total variational distance between $P^{\star}_{i+1,\pi}$ and

$P^{\pi}_{i+1}(\mathcal{D}_N)$:

$$\begin{aligned}
\left|J^{\pi}(\mathcal{M}^{\star}) - J^{\pi}_{\text{Bayes}}(P_{\Theta}(\mathcal{D}_N))\right| &\leq (r_{\max} - r_{\min}) \cdot \sum_{i=0}^{\infty} \gamma^i \text{TV}\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\mathcal{D}_N)\right), \\
&= \frac{r_{\max} - r_{\min}}{1-\gamma} \cdot \sum_{i=0}^{\infty} (1-\gamma)\gamma^i \text{TV}\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\mathcal{D}_N)\right), \\
&= \frac{r_{\max} - r_{\min}}{1-\gamma} \cdot \mathbb{E}_{i\sim\mathcal{G}(\gamma)}\left[\text{TV}\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\mathcal{D}_N)\right)\right], \\
&= \mathcal{R}_{\max} \cdot \mathbb{E}_{i\sim\mathcal{G}(\gamma)}\left[\text{TV}\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\mathcal{D}_N)\right)\right],
\end{aligned}$$

where $\mathcal{G}(\gamma)$ is the geometric distribution. Finally, substituting into Ineq. 13 yields our desired result:

$$\begin{aligned}
\text{Regret}(\mathcal{M}^{\star}, \mathcal{D}_N) &\leq 2 \sup_{\pi} \left|(\mathcal{R}_{\max} \cdot \mathbb{E}_{i\sim\mathcal{G}(\gamma)}\left[\text{TV}\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\mathcal{D}_N)\right)\right]\right|, \\
&= 2\mathcal{R}_{\max} \cdot \sup_{\pi} \mathbb{E}_{i\sim\mathcal{G}(\gamma)}\left[\text{TV}\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\mathcal{D}_N)\right)\right].
\end{aligned}$$

□

We remark that Lemma 2 holds for any general reward-transition model given bounded rewards. The bound in Ineq. 12 proves the true regret is governed by the geometric average of TV distances: $\mathbb{E}_{i\sim\mathcal{G}(\gamma)}\left[\text{TV}\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\mathcal{D}_N)\right)\right]$. As each term $\text{TV}\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\mathcal{D}_N)\right)$ measures the distance between the true and predictive distributions over history h_i of length i, the discounting factor γ determines how much long term histories contribute to regret.

Intuitively, the more mass the posterior places close to the true value $\theta^{\star} \in \Theta^{\star}$, the smaller each TV distance becomes, with regret tending to zero for $P_{i+1,\pi}(\mathcal{D}_N) \approx P^{\star}_{i+1,\pi} \implies \text{TV}\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\mathcal{D}_N)\right) \approx 0$. Conversely, a strong but highly incorrect prior will concentrate mass around MDPs whose dynamics oppose the true dynamics, yielding $\text{TV}\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\mathcal{D}_N)\right) \approx 1$ for all i, achieving the highest possible regret: $\mathcal{R}_{\max} \coloneqq {}^{(r_{\max}-r_{\min})}/_{1-\gamma}$. The resulting Bayes-optimal policy would choose actions that encourage negative reward-seeking behaviour, being farthest from optimal in terms of expected returns.

Using the Bretagnolle-Huber inequality (see Section D.1), we now relate the sum of discounted TV distances to a sum of KL divergences, allowing us to control the expected regret using the PIL.

Theorem 1. *Using the PIL: $\mathcal{I}^{\pi}_N$, the true regret is bounded as:*

$$\text{Regret}(\mathcal{M}^{\star}, \mathcal{D}_N) \leq 2\mathcal{R}_{\max} \cdot \sup_{\pi} \sqrt{1 - \exp\left(-\frac{\mathcal{I}^{\pi}_N}{1-\gamma}\right)}$$

Proof. Starting with the bounded derived in Ineq. 12 of Lemma 2, we apply the Bretagnolle-Huber inequality (Bretagnolle & Huber, 1978) (see Section D.1) to bound the TV distance terms using the KL divergence:

$$\begin{aligned}
\text{Regret}(\mathcal{M}^{\star}, \mathcal{D}_N) &\leq 2\mathcal{R}_{\max} \cdot \sup_{\pi} \mathbb{E}_{i\sim\mathcal{G}(\gamma)}\left[\text{TV}\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\mathcal{D}_N)\right)\right], \\
&\leq 2\mathcal{R}_{\max} \cdot \sup_{\pi} \mathbb{E}_{i\sim\mathcal{G}(\gamma)}\left[\sqrt{1 - \exp\left(-\text{KL}\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\mathcal{D}_N)\right)\right)}\right]. \qquad (14)
\end{aligned}$$

We make two observations. Firstly, as the KL divergence is convex in its second argument and $P_{i+1,\pi}(\mathcal{D}_N) = \mathbb{E}_{\theta\sim P_{\Theta}(\mathcal{D}_N)}\left[P^{\pi}_{i+1}(\theta)\right]$, we can bound each KL divergence term using Jensen's inequality as:

$$\text{KL}\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\mathcal{D}_N)\right) \leq \mathbb{E}_{\theta\sim P_{\Theta}(\mathcal{D}_N)}\left[\text{KL}\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\theta)\right)\right].$$

Secondly, as the function $f(x) = \sqrt{1 - \exp(-x)}$ is monotonically increasing in x, it follows that $f(x) \leq f(x')$ for any $x \leq x'$, hence:

$$\sqrt{1 - \exp\left(-\mathrm{KL}\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\mathcal{D}_N)\right)\right)} \leq \sqrt{1 - \exp\left(-\mathbb{E}_{\theta \sim P_{\Theta}(\mathcal{D}_N)}\left[\mathrm{KL}\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\theta)\right)\right]\right)}$$

Applying this bound to Ineq. 14 yields:

$$\begin{aligned}
&\mathbb{E}_{i \sim \mathcal{G}(\gamma)}\left[\sqrt{1 - \exp\left(-\mathrm{KL}\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\mathcal{D}_N)\right)\right)}\right] \\
&\qquad \leq \mathbb{E}_{i \sim \mathcal{G}(\gamma)}\left[\sqrt{1 - \exp\left(-\mathbb{E}_{\theta \sim P_{\Theta}(\mathcal{D}_N)}\left[\mathrm{KL}\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\theta)\right)\right]\right)}\right].
\end{aligned}$$

As the function $f(x) = \sqrt{1 - \exp(-x)}$ is concave in x, we can apply Jensen's inequality, yielding:

$$\begin{aligned}
&\mathbb{E}_{i \sim \mathcal{G}(\gamma)}\left[\sqrt{1 - \exp\left(-\mathrm{KL}\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\mathcal{D}_N)\right)\right)}\right] \\
&\qquad \leq \sqrt{1 - \exp\left(-\mathbb{E}_{i \sim \mathcal{G}(\gamma)}\left[\mathbb{E}_{\theta \sim P_{\Theta}(\mathcal{D}_N)}\left[\mathrm{KL}\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\theta)\right)\right]\right]\right)}. \qquad (15)
\end{aligned}$$

Examining the KL divergence term:

$$\begin{aligned}
&\mathrm{KL}\left(P^{\star}_{i+1,\pi} \| P^{\pi}_{i+1}(\theta)\right) \\
&\quad = \mathbb{E}_{h_{i+1} \sim P^{\star}_{i+1,\pi}}\left[\log\left(\frac{d_0(s_0)\prod_{j=0}^{i}\pi(a_j|h_j)p^{\star}(r_j|s_j,a_j)p^{\star}(s_{j+1}|s_j,a_j)}{d_0(s_0)\prod_{j=0}^{i}\pi(a_j|h_j)p(r_j|s_j,a_j,\theta)p(s_{j+1}|s_j,a_j,\theta)}\right)\right], \\
&\quad = \mathbb{E}_{h_{i+1} \sim P^{\star}_{i+1,\pi}}\left[\log\left(\frac{\prod_{j=0}^{i}p^{\star}(r_j|s_j,a_j)p^{\star}(s_{j+1}|s_j,a_j)}{\prod_{j=0}^{i}p(r_j|s_j,a_j,\theta)p(s_{j+1}|s_j,a_j,\theta)}\right)\right], \\
&\quad = \mathbb{E}_{h_{i+1} \sim P^{\star}_{i+1,\pi}}\Bigg[\sum_{j=0}^{i}\Big(\log p^{\star}(r_j|s_j,a_j) - \log p(r_j|s_j,a_j,\theta) \\
&\qquad\qquad + \log p^{\star}(s_{j+1}|s_j,a_j) - \log p(s_{j+1}|s_j,a_j,\theta)\Big)\Bigg], \\
&\quad = \sum_{j=0}^{i}\mathbb{E}_{h_j \sim P^{\star}_{j,\pi}}\Bigg[\Big(\log p^{\star}(r_j|s_j,a_j) - \log p(r_j|s_j,a_j,\theta) \\
&\qquad\qquad + \log p^{\star}(s_{j+1}|s_j,a_j) - \log p(s_{j+1}|s_j,a_j,\theta)\Big)\Bigg], \\
&\quad = \sum_{j=0}^{i}\mathbb{E}_{s_j,a_j \sim P^{\star}_{j,\pi}}\Bigg[\mathbb{E}_{r_j,s_{j+1} \sim P^{\star}_{R,S}(s_j,a_j)}\Bigg[\Big(\log p^{\star}(r_j|s_j,a_j) - \log p(r_j|s_j,a_j,\theta) \\
&\qquad\qquad + \log p^{\star}(s_{j+1}|s_j,a_j) - \log p(s_{j+1}|s_j,a_j,\theta)\Big)\Bigg]\Bigg], \\
&\quad = \sum_{j=0}^{i}\mathbb{E}_{s_j,a_j \sim P^{\star}_{j,\pi}}\Bigg[\mathbb{E}_{r_j,s_{j+1} \sim P^{\star}_{R,S}(s_j,a_j)}\Bigg[\Big(\log p^{\star}(r_j,s_{j+1}|s_j,a_j) \\
&\qquad\qquad - \log p(r_j,s_{j+1}|s_j,a_j,\theta)\Big)\Bigg]\Bigg], \\
&\quad = \sum_{j=0}^{i}\mathbb{E}_{s,a \sim P^{\star}_{j,\pi}}\left[\mathrm{KL}(P^{\star}_{R,S}(s,a) \| P_{R,S}(s,a,\theta))\right],
\end{aligned}$$

hence:

$$
\begin{aligned}
&\mathbb{E}_{i\sim\mathcal{G}(\gamma)}\left[\mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\mathrm{KL}\left(P^\star_{i+1,\pi}\|P^\pi_{i+1}(\theta)\right)\right]\right] \\
&= \mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\mathbb{E}_{i\sim\mathcal{G}(\gamma)}\left[\sum_{j=0}^{i}\mathbb{E}_{s,a\sim P^\star_{j,\pi}}\left[\mathrm{KL}(P^\star_{R,S}(s,a)\|P_{R,S}(s,a,\theta))\right]\right]\right], \\
&= \mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\sum_{i=0}^{\infty}(1-\gamma)\gamma^i\sum_{j=0}^{i}\mathbb{E}_{s,a\sim P^\star_{j,\pi}}\left[\mathrm{KL}(P^\star_{R,S}(s,a)\|P_{R,S}(s,a,\theta))\right]\right], \\
&= \mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\sum_{i=0}^{\infty}(1-\gamma)\gamma^i(i+1)\sum_{j=0}^{i}\frac{1}{i+1}\mathbb{E}_{s,a\sim P^\star_{j,\pi}}\left[\mathrm{KL}(P^\star_{R,S}(s,a)\|P_{R,S}(s,a,\theta))\right]\right], \\
&= \frac{1}{1-\gamma}\mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\sum_{i=0}^{\infty}(1-\gamma)^2\gamma^i(i+1)\mathbb{E}_{j\sim\mathcal{U}_i}\left[\mathbb{E}_{s,a\sim P^\star_{j,\pi}}\left[\mathrm{KL}(P^\star_{R,S}(s,a)\|P_{R,S}(s,a,\theta))\right]\right]\right], \\
&= \frac{1}{1-\gamma}\mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\mathbb{E}_{i\sim\mathcal{AG}(\gamma)}\left[\mathbb{E}_{j\sim\mathcal{U}_i}\left[\mathbb{E}_{s,a\sim P^\star_{j,\pi}}\left[\mathrm{KL}(P^\star_{R,S}(s,a)\|P_{R,S}(s,a,\theta))\right]\right]\right]\right]. \qquad (16)
\end{aligned}
$$

Now, as $\rho^\star_\pi = \mathbb{E}_{i\sim\mathcal{AG}(\gamma)}\left[\mathbb{E}_{j\sim\mathcal{U}_i}\left[P^\star_{j,\pi}\right]\right]$ is the arithemetico-geometric ergodic state-action distribution, we can simplify Eq. (16) to yield:

$$
\begin{aligned}
&\mathbb{E}_{i\sim\mathcal{G}(\gamma)}\left[\mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\mathrm{KL}\left(P^\star_{i+1,\pi}\|P^\pi_{i+1}(\theta)\right)\right]\right] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{\mathcal{D}_N\sim P_{\text{Data}}}\left[\mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\mathbb{E}_{s,a\sim\rho^\star_\pi}\left[\mathrm{KL}(P^\star_{R,S}(s,a)\|P_{R,S}(s,a,\theta))\right]\right]\right], \\
&= \frac{1}{1-\gamma}\mathbb{E}_{\mathcal{D}_N\sim P_{\text{Data}}}\left[\mathbb{E}_{s,a\sim\rho^\star_\pi}\left[\mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\mathrm{KL}(P^\star_{R,S}(s,a)\|P_{R,S}(s,a,\theta))\right]\right]\right], \\
&= \frac{1}{1-\gamma}\mathcal{I}^\pi_N,
\end{aligned}
$$

hence, substituting into Ineq. 15, we obtain:

$$
\mathbb{E}_{i\sim\mathcal{G}(\gamma)}\left[\sqrt{1-\exp\left(-\mathrm{KL}\left(P^\star_{i+1,\pi}\|P^\pi_{i+1}(\mathcal{D}_N)\right)\right)}\right] \leq \sqrt{1-\exp\left(-\frac{\mathcal{I}^\pi_N}{1-\gamma}\right)}.
$$

Finally, substituting into Eq. (14) yields our desired result:

$$
\begin{aligned}
\mathrm{Regret}(\mathcal{M}^\star,\mathcal{D}_N) &\leq 2\mathcal{R}_{\max}\cdot\sup_\pi\left|\mathbb{E}_{i\sim\mathcal{G}(\gamma)}\left[\sqrt{1-\exp\left(-\mathrm{KL}\left(P^\star_{i+1,\pi}\|P^\pi_{i+1}(\mathcal{D}_N)\right)\right)}\right]\right|, \\
&\leq 2\mathcal{R}_{\max}\cdot\sup_\pi\sqrt{1-\exp\left(-\frac{\mathcal{I}^\pi_N}{1-\gamma}\right)}.
\end{aligned}
$$

□

The PIL has an intuitive information-geometric interpretation: the inner expectation $\mathbb{E}_{s,a\sim\rho^\star_\pi}\left[\mathrm{KL}\left(P^\star_{R,S}(s,a)\|P_{R,S}(s,a,\theta)\right)\right]$ measures the distance between the model and the true distribution in terms of the information lost when approximating $P^\star_{R,S}(s,a)$ with $P_{R,S}(s,a,\theta)$, averaged across all states. The PIL thus measures how close the posterior's belief is to the truth according to the average information lost under the posterior expectation. We observe that via Jensen's inequality, the PIL is an upper bound on the classic KL risk (sometimes known as expected relative entropy) from Bayesian asymptotics and regret analysis (Aitchison, 1975; Clarke & Barron, 1990; Komaki, 1996; Hartigan, 1998; Barron, 1988; 1999; Yang & Barron, 1999; van der Vaart & van Zanten, 2011; Aslan, 2006; Alaa & van der Schaar, 2018; Bilodeau et al., 2021).

By substituting in our definition of the Gaussian world model, we now find a convenient form for the PIL:

Proposition 1. *Using the Gaussian world model in Eq. (3), it follows:*

$$\mathcal{I}_N^\pi = \mathcal{E}(\mathcal{D}_N, \mathcal{M}^\star) + \mathcal{V}(\mathcal{D}_N).$$

Proof. We substitute the Gaussian world model into the KL divergence to yield:

$$\begin{aligned}
&\mathrm{KL}\left(P^\star_{R,S}(s,a) \| P_{R,S}(s,a,\theta)\right) \\
&= \mathbb{E}_{r,s' \sim P^\star_{R,S}(s,a)} \left[\log \left(\exp\left(-\frac{\|r^\star(s,a) - r\|_2^2}{2\sigma_r^2} \right) \exp\left(-\frac{\|s'^\star(s,a) - s'\|_2^2}{2\sigma_s^2} \right) \right) \right] \\
&\quad - \mathbb{E}_{r,s' \sim P^\star_{R,S}(s,a)} \left[\log \left(\exp\left(-\frac{\|r_\theta(s,a) - r\|_2^2}{2\sigma_r^2} \right) \exp\left(-\frac{\|s'_\theta(s,a) - s'\|_2^2}{2\sigma_s^2} \right) \right) \right], \\
&= \mathbb{E}_{r,s' \sim P^\star_{R,S}(s,a)} \left[\frac{\|r_\theta(s,a) - r\|_2^2 - \|r^\star(s,a) - r\|_2^2}{2\sigma_r^2} \right. \\
&\qquad\qquad \left. + \frac{\|s'_\theta(s,a) - s'\|_2^2 - \|s'^\star(s,a) - s'\|_2^2}{2\sigma_s^2} \right], \\
&= \mathbb{E}_{r,s' \sim P^\star_{R,S}(s,a)} \left[\frac{r_\theta(s,a)^2 - 2r r_\theta(s,a) - r^\star(s,a)^2 + 2r r^\star(s,a)}{2\sigma_r^2} \right. \\
&\qquad\qquad \left. + \frac{\|s'_\theta(s,a)\|_2^2 - 2s'^\top s'_\theta(s,a) - \|s'^\star(s,a)\|_2^2 + 2s'^\top s'^\star(s,a)}{2\sigma_s^2} \right], \\
&= \frac{r_\theta(s,a)^2 - 2r^\star(s,a) r_\theta(s,a) - r^\star(s,a)^2 + 2r^\star(s,a)^2}{2\sigma_r^2} \\
&\qquad + \frac{\|s'_\theta(s,a)\|_2^2 - 2s'^\star(s,a)^\top s'_\theta(s,a) - \|s'^\star(s,a)\|_2^2 + 2\|s'^\star(s,a)\|_2^2}{2\sigma_s^2}, \\
&= \frac{r_\theta(s,a)^2 - 2r^\star(s,a) r_\theta(s,a) + r^\star(s,a)^2}{2\sigma_r^2} \\
&\qquad + \frac{\|s'_\theta(s,a)\|_2^2 - 2s'^\star(s,a)^\top s'_\theta(s,a) + \|s'^\star(s,a)\|_2^2}{2\sigma_s^2}.
\end{aligned}$$

Now, taking expectations with respect to the posterior:

$$\begin{aligned}
&\mathbb{E}_{\theta \sim P_\Theta(\mathcal{D}_N)} \left[\mathrm{KL}\left(P^\star_{R,S}(s,a) \| P_{R,S}(s,a,\theta)\right) \right] \\
&\qquad = \mathbb{E}_{\theta \sim P_\Theta(\mathcal{D}_N)} \left[\frac{r_\theta(s,a)^2 - 2r^\star(s,a) r_\theta(s,a) + r^\star(s,a)^2}{2\sigma_r^2} \right] \\
&\qquad\quad + \mathbb{E}_{\theta \sim P_\Theta(\mathcal{D}_N)} \left[\frac{\|s'_\theta(s,a)\|_2^2 - 2s'^\star(s,a)^\top s'_\theta(s,a) + \|s'^\star(s,a)\|_2^2}{2\sigma_s^2} \right], \\
&\qquad = \frac{\mathbb{E}_{\theta \sim P_\Theta(\mathcal{D}_N)} \left[r_\theta(s,a)^2 \right] - 2r^\star(s,a) r(s,a,\mathcal{D}_N) + r^\star(s,a)^2}{2\sigma_r^2} \\
&\qquad\quad + \frac{\mathbb{E}_{\theta \sim P_\Theta(\mathcal{D}_N)} \left[\|s'_\theta(s,a)\|_2^2 \right] - 2s'^\star(s,a)^\top s'(s,a,\mathcal{D}_N) + \|s'^\star(s,a)\|_2^2}{2\sigma_s^2}.
\end{aligned}$$

Now, we use the variance identity for both the reward and state functions: $\mathbb{E}_{\theta \sim P_\Theta(\mathcal{D}_N)} \left[r_\theta(s,a)^2 \right] = \mathbb{V}_{\theta \sim P_\Theta(\mathcal{D}_N)} \left[r_\theta(s,a) \right] + r(s,a,\mathcal{D}_N)^2$ and $\mathbb{E}_{\theta \sim P_\Theta(\mathcal{D}_N)} \left[\|s'_\theta(s,a)\|_2^2 \right] = \mathbb{V}_{\theta \sim P_\Theta(\mathcal{D}_N)} \left[\|s'_\theta(s,a)\|_2 \right] +$

$\|s'(s,a,\mathcal{D}_N)\|_2^2$ yielding:

$$
\begin{aligned}
&\mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\mathrm{KL}\left(P^\star_{R,S}(s,a)\|P_{R,S}(s,a,\theta)\right)\right] \\
&\quad= \frac{\mathbb{V}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[r_\theta(s,a)\right] + r(s,a,\mathcal{D}_N)^2 - 2r^\star(s,a)r(s,a,\mathcal{D}_N) + r^\star(s,a)^2}{2\sigma_r^2} \\
&\qquad+ \frac{\mathbb{V}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\|s'_\theta(s,a)\|_2\right] + \|s'(s,a,\mathcal{D}_N)\|_2^2 - 2s^{\star\prime}(s,a)^\top s'(s,a,\mathcal{D}_N) + \|s^{\star\prime}(s,a)\|_2^2}{2\sigma_s^2}, \\
&\quad= \frac{\mathbb{V}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[r_\theta(s,a)\right] + (r(s,a,\mathcal{D}_N)^2 - r^\star(s,a))^2}{2\sigma_r^2} \\
&\qquad+ \frac{\mathbb{V}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\|s'_\theta(s,a)\|_2\right] + \|s'(s,a,\mathcal{D}_N) - s^{\star\prime}(s,a)\|_2^2}{2\sigma_s^2}, \\
&\quad= \frac{\mathbb{V}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[r_\theta(s,a)\right]}{2\sigma_r^2} + \frac{\mathbb{V}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\|s'_\theta(s,a)\|_2\right]}{2\sigma_s^2} \\
&\qquad+ \frac{(r(s,a,\mathcal{D}_N) - r^\star(s,a))^2}{2\sigma_r^2} + \frac{\|s'(s,a,\mathcal{D}_N) - s^{\star\prime}(s,a)\|_2^2}{2\sigma_s^2}, \\
&\quad= \mathcal{E}(\mathcal{D}_N,\mathcal{M}^\star) + \mathcal{V}(\mathcal{D}_N),
\end{aligned}
$$

and hence:

$$
\begin{aligned}
\mathcal{I}_N^\pi &= \mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\mathrm{KL}\left(P^\star_{R,S}(s,a)\|P_{R,S}(s,a,\theta)\right)\right], \\
&= \mathcal{E}(\mathcal{D}_N,\mathcal{M}^\star) + \mathcal{V}(\mathcal{D}_N).
\end{aligned}
$$

□

D.3 Proof of Theorem 2

We first introduce some simplifying notation for the expected cross entropy, log likelihood and corresponding gradients and Hessian:

$$
\begin{aligned}
\ell(\theta) &\coloneqq \mathbb{E}_{s,a\sim\rho^\star_\pi, r,s'\sim P^\star_{R,S}(s,a)}\left[\log p(r,s'|s,a,\theta)\right], \\
\ell^\star &\coloneqq \max_{\theta\in\Theta}\ell(\theta) = \mathbb{E}_{s,a\sim\rho^\star_\pi, r,s'\sim P^\star_{R,S}(s,a)}\left[\log p^\star(r,s'|s,a)\right], \\
\ell_N(\theta) &\coloneqq \frac{1}{N}\sum_{i=0}^{N-1}\log p(r_i,s'_i|s_i,a_i,\theta), \\
g^\star_{i,N} &\coloneqq \sqrt{N}\nabla_\theta\ell_N(\theta)\big|_{\theta=\theta^\star_i}, \\
H^\star_i &\coloneqq \nabla^2_\theta\ell(\theta)\big|_{\theta=\theta^\star_i}.
\end{aligned}
$$

We now introduce key regularity assumptions for our parametric model that are required to derive the convergence rate for PIL. They are relatively mild and commonplace in the asymptotic statistics literature (Le Cam, 1953; Barron, 1988; Clarke & Barron, 1990; Komaki, 1996; Hartigan, 1998; Barron, 1999; Aslan, 2006).

Assumption 1. *We assume that:*

I. There exists at least one parametrisation that corresponds to the true environment dynamics with:

$$\left|\mathbb{E}_{s,a\sim\rho^\star_\pi, r,s'\sim P^\star_{R,S}(s,a)}\left[\log p^\star(r,s'|s,a)\right]\right| < \infty$$

and $|\ell^\star - \ell(\theta)|$ is bounded P_Θ-almost surely.

II. $\ell_N(\theta)$ and $\ell(\theta)$ are C^2-continuous in θ.

III. There are $K < \infty$ maximising points $\theta_i^\star$:

$$\{\theta_1^\star, \theta_2^\star, \dots \theta_K^\star\} = \arg\max_{\theta \in \Theta} \ell(\theta).$$

For each maximiser $\theta_i^\star$, there exists a small region $\Theta_i^\star := \{\theta \in \Theta | \|\theta_i^\star - \theta\| \leq \epsilon\}$ for some $\epsilon > 0$ such that $\theta_i^\star$ is the unique maximiser in $\Theta_i^\star$, $\theta_i^\star$ is in the interior of $\Theta_i^\star$, $\nabla_\theta^2 \ell(\theta_i^\star)$ is negative definite, invertible and the regions are disjoint: $\bigcap_{i=1}^K \Theta_i^\star = \varnothing$.

IV. The prior $p(\theta)$ is Lipschitz continuous in θ with support over Θ.

V. The sampling regime ensures that the strong law of large numbers holds for all maximisers $\theta_i^\star$ for the Hessian, and uniformly for $\theta \in \Theta$ for the likelihood, that is:

$$\ell_N(\theta) \xrightarrow{\text{Unif. a.s.}} \ell(\theta), \quad \nabla_\theta^2 \ell_N(\theta_i^\star) \xrightarrow{\text{a.s.}} \nabla_\theta^2 \ell(\theta_i^\star).$$

The central limit theorem applies to the gradient at each $\theta_i^\star$, that is:

$$\sqrt{N} \nabla_\theta \ell_N(\theta_i^\star) \xrightarrow{d} \mathcal{N}(0, \Sigma_i^g),$$

where $\Sigma_i^g = \mathbb{E}_{s,a \sim \rho_\pi^\star, r, s' \sim P^\star_{R,S}(s,a)} \left[\nabla_\theta \log p(r, s' | s, a, \theta_i^\star) \nabla_\theta \log p(r, s' | s, a, \theta_i^\star)^\top \right]$ with $\|\Sigma_i^g\| < \infty$.

Assumption 1i is our strictest assumption and is included for ease of exposition. We generalise our theory in Section D.4 to relax this assumption and also account for incomplete Bayes-optimal policy learning. Assumption 1ii ensures that a second order Taylor series expansion can be applied to obtain an asymptotic expansion around the maximising points. Assumption 1iii is much more general than most settings, which only consider problems with a single maximiser. The invertibility of the matrix can easily be guaranteed in Bayesian methods by the use of a prior that can re-condition a low rank matrix that may results from linearly dependent data. Assumption 1iv ensures that the prior places sufficient mass on the true parametrisation. The sampling and model would need to be very irregular for Assumption 1v not to hold; stochastic optimisation methods used to find statistics like the MAP will fail if this assumption did not hold. Assumption 1v holds automatically if sampling is either i.i.d. from $s, a \sim \rho_\pi^\star$ (see e.g. Bass (2013)) or from an aperiodic and irreducible Markov chain with stationary distribution $\rho_\pi^\star$ (see e.g. Roberts & Rosenthal (2004)). In both sample regimes, noting that $\mathbb{E}_{s,a \sim \rho_\pi^\star, r, s' \sim P^\star_{R,S}(s,a)} \left[\nabla_\theta \log p(r, s' | s, a, \theta_i^\star) \right] = 0$, it's clear the (long run) covariance of $\nabla_\theta \log p(r, s' | s, a, \theta_i^\star)$ is Σ_i^g.

Our first lemma borrows techniques from Vaart (1998, Chapter 10). This approach is similar to asymptotic integral expansion approaches that apply Laplace's method (Lindley, 1980; Tierney & Kadane, 1986; Tierney et al., 1989; Kass et al., 1990) except we expand around the global maximising values of $\ell(\theta)$ rather than the maximising values of the likelihood $\ell_N(\theta)$ to obtain an asymptotic expression for the posterior:

Lemma 3. *Under Assumption 1 and using the notation introduced at the start of Section D.3:*

$$\frac{\int_{\Theta_i^\star} \left(\ell^\star - \ell(\theta) \exp\left(N \ell_N(\theta) \right) \right) p(\theta) d\theta}{\int_{\Theta_i^\star} \exp\left(N \ell_N(\theta) \right) p(\theta) d\theta} = \mathcal{O}\left(\frac{d - {g_{i,N}^\star}^\top {H_i^\star}^{-1} g_{i,N}^\star}{N} \right),$$

almost surely.

Proof. We start by applying the transformation of variables $\theta' = f(\theta) := \sqrt{N}(\theta - \theta_i^\star)$ to integrals in the numerator and denominator with:

$$\theta = f^{-1}(\theta') = \theta_i^\star + \frac{1}{\sqrt{N}} \theta', \quad \left| \det \nabla_\theta f^{-1}(\theta') \right| = N^{-\frac{d}{2}}, \quad \Theta' := f(\Theta_i^\star),$$

yielding:

$$\frac{\int_{\Theta_i^\star}(\ell^\star - \ell(\theta))\exp\left(N\ell_N(\theta)\right)p(\theta)d\theta}{\int_{\Theta^\star}\exp\left(N\ell_N(\theta)\right)p(\theta)d\theta} = \frac{\int_{\Theta'}\left(\ell^\star - \ell\left(\theta = \theta^\star + \frac{1}{\sqrt{N}}\theta'\right)\right)\exp\left(N\ell_N\left(\theta = \theta_i^\star + \frac{1}{\sqrt{N}}\theta'\right)\right)p'(\theta')d\theta'}{\int_{\Theta'}\exp\left(N\ell_N\left(\theta = \theta_i^\star + \frac{1}{\sqrt{N}}\theta'\right)\right)p'(\theta')d\theta'}, \tag{17}$$

where $p'(\theta') \coloneqq p\left(\theta = \theta_i^\star + \frac{1}{\sqrt{N}}\theta'\right)$. Now and making a Taylor series expansion of $\ell(\theta)$ about $\theta_i^\star$:

$$\begin{aligned}\ell(\theta) &= \ell^\star + \underbrace{\nabla_\theta \ell(\theta_i^\star)}^{\top}(\theta - \theta_i^\star) + (\theta - \theta_i^\star)^\top H_i^\star(\theta - \theta_i^\star) + \mathcal{O}\left(\|\theta - \theta_i^\star\|^3\right), \\ &= \ell^\star + (\theta - \theta_i^\star)^\top H_i^\star(\theta - \theta_i^\star) + \mathcal{O}\left(\|\theta - \theta_i^\star\|^3\right),\end{aligned}$$

hence:

$$\ell\left(\theta = \theta_i^\star + \frac{1}{\sqrt{N}}\theta'\right) = \ell^\star + \frac{1}{N}\theta'^\top H_i^\star\theta' + \mathcal{O}\left(N^{-\frac{3}{2}}\right).$$

Using the notation $H_N^\star \coloneqq \nabla_\theta^2 \ell_N(\theta)\big|_{\theta=\theta_i^\star}$ and making a Taylor series expansion of $\ell_N(\theta)$ about $\theta_i^\star$:

$$\ell_N(\theta) = \ell_N(\theta_i^\star) + \nabla_\theta \ell_N(\theta_i^\star)^\top(\theta - \theta_i^\star) + (\theta - \theta_i^\star)^\top \nabla_\theta^2 \ell_N(\theta_i^\star)(\theta - \theta_i^\star) + \mathcal{O}\left(\|\theta - \theta_i^\star\|^3\right),$$

hence:

$$N\ell_N\left(\theta = \theta_i^\star + \frac{1}{\sqrt{N}}\theta'\right) = N\ell_N(\theta_i^\star) + \sqrt{N}\nabla_\theta \ell_N(\theta_i^\star)^\top\theta' + \theta'^\top\nabla_\theta^2\ell_N(\theta_i^\star)\theta' + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right).$$

Substituting into Eq. (17) yields:

$$\begin{aligned}&\frac{\int_{\Theta^\star}(\ell^\star - \ell(\theta))\exp\left(N\ell(\theta)\right)p(\theta)d\theta}{\int_{\Theta^\star}\exp\left(N\ell(\theta)\right)p(\theta)d\theta} \\ &= -\frac{\int_{\Theta'}\theta'^\top H_i^\star\theta'\exp\left(N\ell_N(\theta_i^\star) + {g_{i,N}^\star}^\top\theta' + \theta'^\top H_N^\star\theta' + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)\right)p'(\theta')d\theta'}{\int_{\Theta'}\exp\left(N\ell_N(\theta_i^\star) + {g_{i,N}^\star}^\top\theta' + \theta'^\top H_N^\star\theta' + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)\right)p'(\theta')d\theta'}\mathcal{O}\left(\frac{1}{N}\right), \\ &= -\frac{\int_{\Theta'}\theta'^\top H_i^\star\theta'\exp\left({g_{i,N}^\star}^\top\theta' + \theta'^\top H_N^\star\theta' + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)\right)p'(\theta')d\theta'}{\int_{\Theta'}\exp\left(N{g_{i,N}^\star}^\top\theta' + \theta'^\top H_N^\star\theta' + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)\right)p'(\theta')d\theta'}\mathcal{O}\left(\frac{1}{N}\right), \\ &= -\frac{\int_{\Theta'}\theta'^\top H_i^\star\theta'\exp\left({g_{i,N}^\star}^\top\theta' + \theta'^\top H_N^\star\theta'\right)\exp\left(\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)\right)p'(\theta')d\theta'}{\int_{\Theta'}\exp\left(N{g_{i,N}^\star}^\top\theta' + \theta'^\top H_N^\star\theta'\right)\exp\left(\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)\right)p'(\theta')d\theta'}\mathcal{O}\left(\frac{1}{N}\right), \\ &= \mathcal{O}\left(-\frac{1}{N}\frac{\int_{\Theta'}\theta'^\top H_i^\star\theta'\exp\left({g_{i,N}^\star}^\top\theta' + \theta'^\top H_N^\star\theta'\right)p'(\theta')d\theta'}{\int_{\Theta'}\exp\left(N{g_{i,N}^\star}^\top\theta' + \theta'^\top H_N^\star\theta'\right)p'(\theta')d\theta'}\right)\end{aligned} \tag{18}$$

where we have multiplied top and bottom by $\exp\left(-N\ell_N(\theta_i^\star)\right)$ to derive the second equality and used the fact that $0 < \exp\left(\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)\right) < \infty$ to derive the final line. Now, as the prior is Lipschitz, we make a Taylor series expansion about $\theta_i^\star$:

$$p(\theta) = p(\theta_i^\star) + \mathcal{O}\left(\|\theta - \theta_i^\star\|\right),$$

hence:

$$p'(\theta') = p\left(\theta = \theta_i^\star + \frac{1}{\sqrt{N}}\theta'\right) = p(\theta_i^\star) + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right).$$

This allows us to find an asymptotic expression for Eq. (18):

$$\begin{aligned}
&\frac{\int_{\Theta'} \theta'^\top H_i^\star \theta' \exp\left({g_{i,N}^\star}^\top \theta' + \theta'^\top H_N^\star \theta'\right) p'(\theta') d\theta'}{\int_{\Theta'} \exp\left({g_{i,N}^\star}^\top \theta' + \theta'^\top H_N^\star \theta'\right) p'(\theta') d\theta'} \\
&= \frac{\int_{\Theta'} \theta'^\top H_i^\star \theta' \exp\left({g_{i,N}^\star}^\top \theta' + \theta'^\top H_N^\star \theta'\right) p'(\theta_i^\star) d\theta' \left(1 + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)\right)}{\int_{\Theta'} \exp\left({g_{i,N}^\star}^\top \theta' + \theta'^\top H_N^\star \theta'\right) p'(\theta_i^\star) d\theta' \left(1 + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)\right)} \\
&= \frac{\int_{\Theta'} \theta'^\top H_i^\star \theta' \exp\left({g_{i,N}^\star}^\top \theta' + \theta'^\top H_N^\star \theta'\right) p'(\theta_i^\star) d\theta'}{\int_{\Theta'} \exp\left({g_{i,N}^\star}^\top \theta' + \theta'^\top H_N^\star \theta'\right) p'(\theta_i^\star) d\theta'} \left(1 + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)\right) \\
&= \frac{\int_{\Theta'} \theta'^\top H_i^\star \theta' \exp\left({g_{i,N}^\star}^\top \theta' + \theta'^\top H_N^\star \theta'\right) d\theta'}{\int_{\Theta'} \exp\left({g_{i,N}^\star}^\top \theta' + \theta'^\top H_N^\star \theta'\right) d\theta'} \left(1 + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)\right).
\end{aligned}$$

We re-write the exponential term to recover a quadratic form:

$$\begin{aligned}
&\exp\left({g_{i,N}^\star}^\top \theta' + \theta'^\top H_N^\star \theta'\right) \\
&= \exp\left(\left(\frac{1}{2} {H_N^\star}^{-1} g_{i,N}^\star + \theta'\right)^\top H_N^\star \left(\frac{1}{2} {H_N^\star}^{-1} g_{i,N}^\star + \theta'\right) - \frac{1}{4} {g_{i,N}^\star}^\top {H_N^\star}^{-1} g_{i,N}^\star\right).
\end{aligned}$$

Substituting yields:

$$\begin{aligned}
&\frac{\int_{\Theta'} \theta'^\top H_i^\star \theta' \exp\left({g_{i,N}^\star}^\top \theta' + \theta'^\top H_N^\star \theta'\right) d\theta'}{\int_{\Theta'} \exp\left({g_{i,N}^\star}^\top \theta' + \theta'^\top H_N^\star \theta'\right) d\theta'} \\
&= \frac{\int_{\Theta'} \theta'^\top H_i^\star \theta' \exp\left(\left(\frac{1}{2} {H_N^\star}^{-1} g_{i,N}^\star + \theta'\right)^\top H_N^\star \left(\frac{1}{2} {H_N^\star}^{-1} g_{i,N}^\star + \theta'\right)\right) d\theta'}{\int_{\Theta'} \exp\left(\left(\frac{1}{2} {H_N^\star}^{-1} g_{i,N}^\star + \theta'\right)^\top H_N^\star \left(\frac{1}{2} {H_N^\star}^{-1} g_{i,N}^\star + \theta'\right)\right) d\theta'}.
\end{aligned}$$

In this form, we notice the expectation is that of a Gaussian $\mathcal{N}\left(\mu = -\frac{1}{2} {H_N^\star}^{-1} g_{i,N}^\star, \Sigma = -{H_i^\star}^{-1}\right)$ restricted to Θ'. Noting that in the limit $\Theta' \xrightarrow[N \to \infty]{} \mathbb{R}^d$, hence:

$$\begin{aligned}
&\frac{\int_{\Theta'} \theta'^\top H_i^\star \theta' \exp\left(\left(\theta' + \frac{1}{2} {H_N^\star}^{-1} g_{i,N}^\star\right)^\top H_N^\star \left(\theta' + \frac{1}{2} {H_N^\star}^{-1} g_{i,N}^\star\right)\right) d\theta'}{\int_{\Theta'} \exp\left(\left(\theta' + \frac{1}{2} {H_N^\star}^{-1} g_{i,N}^\star\right)^\top H_N^\star \left(\theta' + \frac{1}{2} {H_N^\star}^{-1} g_{i,N}^\star\right)\right) d\theta'} \\
&= \mathcal{O}\left(\frac{\int_{\mathbb{R}^d} \theta'^\top H_i^\star \theta' \exp\left(\left(\theta' + \frac{1}{2} {H_N^\star}^{-1} g_{i,N}^\star\right)^\top H_N^\star \left(\theta' + \frac{1}{2} {H_N^\star}^{-1} g_{i,N}^\star\right)\right) d\theta'}{\int_{\mathbb{R}^d} \exp\left(\left(\theta' + \frac{1}{2} {H_N^\star}^{-1} g_{i,N}^\star\right)^\top H_N^\star \left(\theta' + \frac{1}{2} {H_N^\star}^{-1} g_{i,N}^\star\right)\right) d\theta'}\right), \\
&= \mathcal{O}\left(\mathbb{E}_{\theta' \sim \mathcal{N}\left(-\frac{1}{2} {H_N^\star}^{-1} g_{i,N}^\star, -{H_N^\star}^{-1}\right)} \left[\theta'^\top H_i^\star \theta'\right]\right). \qquad (19)
\end{aligned}$$

Putting everything together, we have:

$$\begin{aligned}
&\frac{\int_{\Theta_i^\star}\left(\ell^\star - \ell(\theta)\exp\left(N\ell_N(\theta)\right)\right)p(\theta)d\theta}{\int_{\Theta_i^\star}\exp\left(N\ell_N(\theta)\right)p(\theta)d\theta} \\
&\quad= \mathcal{O}\left(-\frac{1}{N}\frac{\int_{\Theta'}\theta'^\top H_i^\star\theta'\exp\left({g_{i,N}^\star}^\top\theta' + \theta'^\top H_N^\star\theta'\right)p'(\theta')d\theta'}{\int_{\Theta'}\exp\left({g_{i,N}^\star}^\top\theta' + \theta'^\top H_N^\star\theta'\right)p'(\theta')d\theta'}\right), \quad \text{Eq. (18)} \\
&\quad= \mathcal{O}\left(-\frac{1}{N}\mathbb{E}_{\theta'\sim\mathcal{N}\left(-\frac{1}{2}{H_N^\star}^{-1}g_{i,N}^\star, -{H_i^\star}^{-1}\right)}\left[\theta'^\top H_i^\star\theta'\right]\right), \quad \text{Eq. (19)}
\end{aligned}$$

Using standard results for the multivariate Gaussian (Petersen & Pedersen, 2012) yields our desired result:

$$\begin{aligned}
-\frac{1}{N}\mathbb{E}_{\theta'\sim\mathcal{N}\left(-\frac{1}{2}{H_N^\star}^{-1}g_{i,N}^\star, -{H_i^\star}^{-1}\right)}\left[\theta'^\top H_i^\star\theta'\right] &= \frac{\mathrm{Tr}\left(H_i^\star{H_N^\star}^{-1}\right) - \frac{1}{4}{g_{i,N}^\star}^\top{H_N^\star}^{-1^\top}H_i^\star{H_N^\star}^{-1}g_{i,N}^\star}{N}, \\
&= \mathcal{O}\left(\frac{\mathrm{Tr}\left(I\right) - {g_{i,N}^\star}^\top H_i^{\star-1}g_{i,N}^\star}{N}\right), \\
&= \mathcal{O}\left(\frac{d - {g_{i,N}^\star}^\top H_i^{\star-1}g_{i,N}^\star}{N}\right),
\end{aligned}$$

almost surely, where we have used the strong law of large numbers on the empirical Hessian from Assumption 1 to derive the second line. □

In our final Lemma, we show that regions that are not close to the maximising points diminish exponentially in posterior probability as N grows large.

Lemma 4. *Under Assumption 1,* $\mathbb{E}_{\mathcal{D}_N\sim P^\star_{Data}}\left[P(\bar{\Theta}|\mathcal{D}_N)\right] = \mathcal{O}\left(\exp(-N)\right)$.

Proof. We start by splitting the posterior expectation into integrals over $\bar{\Theta}$ and $\Theta\setminus\bar{\Theta}$:

$$\begin{aligned}
P(\bar{\Theta}|\mathcal{D}_N) &= \frac{\int_{\bar{\Theta}}\exp\left(N\ell_N(\theta)\right)p(\theta)d\theta}{\int_{\Theta}\exp\left(N\ell_N(\theta)\right)p(\theta)d\theta}, \\
&= \frac{\int_{\bar{\Theta}}\exp\left(N\ell_N(\theta)\right)p(\theta)d\theta}{\int_{\bar{\Theta}}\exp\left(N\ell_N(\theta)\right)p(\theta)d\theta + \int_{\Theta\setminus\bar{\Theta}}\exp\left(N\ell_N(\theta)\right)p(\theta)d\theta}.
\end{aligned}$$

Dividing top and bottom by $\int_{\bar{\Theta}}\exp\left(N\ell_N(\theta)\right)p(\theta)d\theta$:

$$P(\bar{\Theta}|\mathcal{D}_N) = \frac{1}{1 + \frac{\int_{\Theta\setminus\bar{\Theta}}\exp(N\ell_N(\theta))p(\theta)d\theta}{\int_{\bar{\Theta}}\exp(N\ell_N(\theta))p(\theta)d\theta}}.$$

Hence if we can show there exists some $N' < \infty$ and a function $C\exp(cN)$ with positive constants c and C that lower bounds the ratio:

$$C\exp(cN) \leq \frac{\int_{\Theta\setminus\bar{\Theta}}\exp\left(N\ell_N(\theta)\right)p(\theta)d\theta}{\int_{\bar{\Theta}}\exp\left(N\ell_N(\theta)\right)p(\theta)d\theta}$$

almost surely for all $N \geq N'$, then it follows:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_N\sim P^\star_{\text{Data}}}\left[P(\bar{\Theta}|\mathcal{D}_N)\right] &= \mathcal{O}\left(\frac{1}{1+\exp(N)}\right), \\
&= \mathcal{O}\left(\exp(-N)\right).
\end{aligned}$$

From Assumption 1, each $\theta_i^\star$ maximises $\ell(\theta)$ with $\sup_{\theta\in\bar{\Theta}}\ell(\theta') < \ell(\theta_i^\star)$. As $\ell(\theta)$ is continuous, there thus exists a small, closed ball $B(\theta_j^\star, r) \coloneqq \{\theta \| \|\theta_i^\star - \theta\| \leq r\}$ of radius $r > 0$ centred on some $\theta_j^\star$ such that $\sup_{\theta'\in\bar{\Theta}}\ell(\theta') < \min_{\theta''\in B(\theta_j^\star, r)}\ell(\theta'')$. From Assumption 1, the uniform strong

law of large numbers holds with $\ell_N(\theta) \xrightarrow{\text{Unif. } a.s} \ell(\theta)$. By the definition of the limit and continuity of $\ell_N(\theta)$, there thus exists some finite N' such that $\sup_{\theta' \in \bar{\Theta}} \ell_N(\theta') < \min_{\theta'' \in B(\theta_j^\star, \frac{r}{2})} \ell_N(\theta'')$ for all $N \geq N'$ almost surely, where $B(\theta_j^\star, \frac{r}{2})$ is a ball of half radius $\frac{r}{2}$. Noting that $B(\theta_j^\star, \frac{r}{2}) \subset \Theta \setminus \bar{\Theta}$ and $0 \leq \exp\left(N\ell_N(\theta)\right)$, this allows us to lower bound the integral:

$$\begin{aligned}
\int_{\Theta \setminus \bar{\Theta}} \exp\left(N\ell_N(\theta)\right) p(\theta) d\theta &\geq \int_{B(\theta_j^\star, \frac{r}{2})} \exp\left(N\ell_N(\theta)\right) p(\theta) d\theta, \\
&\geq \exp\left(N \min_{\theta'' \in B(\theta_j^\star, \frac{r}{2})} \ell_N(\theta'')\right) \int_{B(\theta_j^\star, \frac{r}{2})} p(\theta) d\theta, \\
&= \exp\left(N \min_{\theta'' \in B(\theta_j^\star, \frac{r}{2})} \ell_N(\theta'')\right) P\left(B\left(\theta_j^\star, \frac{r}{2}\right)\right).
\end{aligned}$$

We can also upper bound the integral:

$$\begin{aligned}
\int_{\bar{\Theta}} \exp\left(N\ell_N(\theta)\right) p(\theta) d\theta &\leq \exp\left(N \sup_{\theta' \in \bar{\Theta}} \ell_N(\theta')\right) \int_{\bar{\Theta}} p(\theta) d\theta, \\
&= \exp\left(N \sup_{\theta' \in \bar{\Theta}} \ell_N(\theta')\right) P\left(\bar{\Theta}\right).
\end{aligned}$$

Using these results, we lower bound the ratio as:

$$\begin{aligned}
\frac{\int_{\Theta \setminus \bar{\Theta}} \exp\left(N\ell_N(\theta)\right) p(\theta) d\theta}{\int_{\bar{\Theta}} \exp\left(N\ell_N(\theta)\right) p(\theta) d\theta} &\geq \frac{\exp\left(N \min_{\theta'' \in B(\theta_j^\star, \frac{r}{2})} \ell_N(\theta'')\right) P\left(B\left(\theta_j^\star, \frac{r}{2}\right)\right)}{\exp\left(N \sup_{\theta' \in \bar{\Theta}} \ell_N(\theta')\right) P\left(\bar{\Theta}\right)}, \\
&= \exp\left(N\left(\min_{\theta'' \in B(\theta_j^\star, \frac{r}{2})} \ell_N(\theta'') - \sup_{\theta' \in \bar{\Theta}} \ell_N(\theta')\right)\right) \frac{P\left(B\left(\theta_j^\star, \frac{r}{2}\right)\right)}{P\left(\bar{\Theta}\right)}.
\end{aligned}$$

Let $\frac{P\left(B\left(\theta_j^\star, \frac{r}{2}\right)\right)}{P\left(\bar{\Theta}\right)} = C > 0$ from Assumption 1. As there exists some N' such that $\min_{\theta'' \in B(\theta_j^\star, \frac{r}{2})} \ell_N(\theta'') > \sup_{\theta' \in \bar{\Theta}} \ell_N(\theta')$ for all $N > N'$, we have shown exists some positive constants $c > 0$ and $C > 0$ such that

$$C \exp\left(cN\right) \leq \frac{\int_{\Theta \setminus \bar{\Theta}} \exp\left(N\ell_N(\theta)\right) p(\theta) d\theta}{\int_{\bar{\Theta}} \exp\left(N\ell_N(\theta)\right) p(\theta) d\theta},$$

for all $N > N'$ almost surely, as required. □

We now present our proof of Theorem 2. Here we split the posterior expectation up into small regions close to maximising points and regions away from maximising. We then apply our two lemmas to each region. Our result then follows by an application the central limit theorem under Assumption 1.

Theorem 2. *Let the data be drawn from the underlying true distribution $\mathcal{D}_N \sim P^\star_{Data}$. Under Assumption 1, there exists some constant $0 < C < \infty$ such that for sufficiently large N:*

$$\mathbb{E}_{\mathcal{D}_N \sim P^\star_{Data}}\left[\text{Regret}(\mathcal{M}^\star, \mathcal{D}_N)\right] \leq 2\mathcal{R}_{\max} \cdot \sqrt{1 - \exp\left(-\frac{Cd}{(1-\gamma)N}\right)}. \tag{20}$$

Proof. Using the notation introduced at the start of Section D.3, we write the PIL as:

$$\begin{aligned}
\mathcal{I}^\pi_N :=& \mathbb{E}_{s,a \sim \rho^\star_\pi}\left[\mathbb{E}_{\theta \sim P_\Theta(\mathcal{D}_N)}\left[\text{KL}\left(P^\star_{R,S}(s,a) \| P_{R,S}(s,a,\theta)\right)\right]\right], \\
=& \mathbb{E}_{\theta \sim P_\Theta(\mathcal{D}_N)}\left[\mathbb{E}_{s,a \sim \rho^\star_\pi, r, s' \sim P^\star_{R,S}(s,a)}\left[\log p(r, s' | s, a, \theta^\star) - \log p(r, s' | s, a, \theta)\right]\right], \\
=& \mathbb{E}_{\theta \sim P_\Theta(\mathcal{D}_N)}\left[\ell^\star - \ell(\theta)\right],
\end{aligned}$$

Under this same notation, we write the posterior density as:

$$p(\theta|\mathcal{D}_N) = \frac{\exp\left(N\ell_N(\theta)\right)p(\theta)}{\int_\Theta \exp\left(N\ell_N(\theta)\right)p(\theta)d\theta}. \tag{21}$$

Now, under Assumption 1, we split the inner expectation into small regions $\Theta_i^\star$ around each maximising point $\theta_i^\star$ and the remainder of the parameter space $\bar{\Theta} \coloneqq \Theta \setminus \bigcup_{i=1}^K \Theta_i^\star$:

$$\mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\ell^\star - \ell(\theta)\right] = \sum_{i=1}^K \int_{\Theta_i^\star} \left(\ell^\star - \ell(\theta)\right)p(\theta|\mathcal{D}_N)d\theta + \int_{\bar{\Theta}} \left(\ell^\star - \ell(\theta)\right)p(\theta|\mathcal{D}_N)d\theta. \tag{22}$$

Using Eq. (21), we now re-write each integral in the summation term of Eq. (22) as:

$$\begin{aligned}
\int_{\Theta_i^\star} \left(\ell^\star - \ell(\theta)\right)p(\theta|\mathcal{D}_N)d\theta &= \frac{\int_{\Theta_i^\star} \left(\ell^\star - \ell(\theta)\right)\exp\left(N\ell_N(\theta)\right)p(\theta)d\theta}{\int_\Theta \exp\left(N\ell_N(\theta)\right)p(\theta)d\theta}, \\
&= \frac{\int_{\Theta_i^\star} \left(\ell^\star - \ell(\theta)\right)\exp\left(N\ell_N(\theta)\right)p(\theta)d\theta}{\int_\Theta \exp\left(N\ell_N(\theta)\right)p(\theta)d\theta} \cdot \frac{\int_{\Theta_i^\star} \exp\left(N\ell_N(\theta)\right)p(\theta)d\theta}{\int_{\Theta_i^\star} \exp\left(N\ell_N(\theta)\right)p(\theta)d\theta}, \\
&= \frac{\int_{\Theta_i^\star} \left(\ell^\star - \ell(\theta)\right)\exp\left(N\ell_N(\theta)\right)p(\theta)d\theta}{\int_{\Theta_i^\star} \exp\left(N\ell_N(\theta)\right)p(\theta)d\theta} \cdot \frac{\int_{\Theta_i^\star} \exp\left(N\ell_N(\theta)\right)p(\theta)d\theta}{\int_\Theta \exp\left(N\ell_N(\theta)\right)p(\theta)d\theta}, \\
&= \frac{\int_{\Theta_i^\star} \left(\ell^\star - \ell(\theta)\right)\exp\left(N\ell_N(\theta)\right)p(\theta)d\theta}{\int_{\Theta_i^\star} \exp\left(N\ell_N(\theta)\right)p(\theta)d\theta} \cdot P(\Theta_i^\star|\mathcal{D}_N), \\
&\leq \frac{\int_{\Theta_i^\star} \left(\ell^\star - \ell(\theta)\right)\exp\left(N\ell_N(\theta)\right)p(\theta)d\theta}{\int_{\Theta_i^\star} \exp\left(N\ell_N(\theta)\right)p(\theta)d\theta}.
\end{aligned} \tag{23}$$

where we have used $0 \leq P(\Theta_i^\star|\mathcal{D}_N) \leq 1$ from Kolmogorov's axioms to bound the final line.

For the last term in Eq. (22), we note that $\ell^\star - \ell(\theta)$ is bounded P_Θ-almost surely from Assumption 1, hence there exists some $\ell^\dagger < \infty$ such that:

$$\begin{aligned}
\int_{\bar{\Theta}} \left(\ell^\star - \ell(\theta)\right)p(\theta|\mathcal{D}_N)d\theta &\leq \int_{\bar{\Theta}} \ell^\dagger p(\theta|\mathcal{D}_N)d\theta, \\
&= \ell^\dagger P(\bar{\Theta}|\mathcal{D}_N).
\end{aligned} \tag{24}$$

Using Ineqs. 23 and 24, we bound Eq. (22) as:

$$\mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\ell^\star - \ell(\theta)\right] \leq \sum_{i=1}^K \frac{\int_{\Theta_i^\star} \left(\ell^\star - \ell(\theta)\right)\exp\left(N\ell_N(\theta)\right)p(\theta)d\theta}{\int_{\Theta_i^\star} \exp\left(N\ell_N(\theta)\right)p(\theta)d\theta} + \ell^\dagger P(\bar{\Theta}|\mathcal{D}_N),$$

and hence the PIL can be bounded as:

$$\mathcal{I}_N^\pi \leq \sum_{i=1}^K \frac{\int_{\Theta_i^\star} \left(\ell^\star - \ell(\theta)\right)\exp\left(N\ell_N(\theta)\right)p(\theta)d\theta}{\int_{\Theta_i^\star} \exp\left(N\ell_N(\theta)\right)p(\theta)d\theta} + \ell^\dagger P(\bar{\Theta}|\mathcal{D}_N).$$

Applying Lemma 3 and Lemma 4 under Assumption 1 yields:

$$\begin{aligned}
\mathcal{I}_N^\pi &= \sum_{i=1}^K \mathcal{O}\left(\frac{d - {g_{i,N}^\star}^\top {H_i^\star}^{-1} g_{i,N}^\star}{N}\right) + \ell^\dagger \mathcal{O}\left(\exp(-N)\right), \\
&= \mathcal{O}\left(\frac{d - \sum_{i=1}^K {g_{i,N}^\star}^\top {H_i^\star}^{-1} g_{i,N}^\star}{N}\right).
\end{aligned} \tag{25}$$

almost surely. As $f(x) := 2\mathcal{R}_{\max} \cdot \sqrt{1-\exp\left(-\frac{x}{(1-\gamma)}\right)}$ is monotonic in x and $\frac{d-\sum_{i=1}^{K} {g_{i,N}^{\star}}^{\top} H_i^{\star -1} g_{i,N}^{\star}}{N} \geq 0$, Eq. (25) implies there exists some positive $0 < C < \infty$ such that:

$$\text{Regret}(\mathcal{M}^{\star}, \mathcal{D}_N) \leq 2\mathcal{R}_{\max} \cdot \sqrt{1-\exp\left(-C\frac{d-\sum_{i=1}^{K} {g_{i,N}^{\star}}^{\top} H_i^{\star -1} g_{i,N}^{\star}}{(1-\gamma)N}\right)},$$

almost surely for large enough N. Under Assumption 1, $g_{i,N}^{\star} \xrightarrow{d} \mathcal{N}(0, \Sigma_i^g)$. As $f(x)$ is also a bounded, continuous function and concave, we can apply the Portmanteau Theorem (see for example Bass (2013, Chapter 21.7)) followed by Jensen's inequality to yield:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_N \sim P_{\text{Data}}^{\star}}\left[\text{Regret}(\mathcal{M}^{\star}, \mathcal{D}_N)\right] \leq & 2\mathcal{R}_{\max} \cdot \mathbb{E}_{g_i \sim \mathcal{N}(0,\Sigma_i^g)}\left[\sqrt{1-\exp\left(-C\frac{d-\sum_{i=1}^{K} {g_i}^{\top} H_i^{\star -1} g_i}{(1-\gamma)N}\right)}\right], \\
\leq & 2\mathcal{R}_{\max} \cdot \sqrt{1-\exp\left(-C\frac{d-\sum_{i=1}^{K} \mathbb{E}_{g_i \sim \mathcal{N}(0,\Sigma_i^g)}\left[g^{\top} H_i^{\star -1} g\right]}{(1-\gamma)N}\right)}, \\
= & 2\mathcal{R}_{\max} \cdot \sqrt{1-\exp\left(-C\frac{d-\sum_{i=1}^{K} \text{Tr}\left(\Sigma_i^g H_i^{\star -1}\right)}{(1-\gamma)N}\right)}. \qquad (26)
\end{aligned}$$

Now, examining the Hessian:

$$\begin{aligned}
H(\theta) &= \nabla_{\theta}^2 \mathbb{E}_{s,a \sim \rho_{\pi}^{\star}, r, s' \sim P_{R,S}^{\star}(s,a)}\left[\log p(r, s'|s, a, \theta)\right] \\
&= \nabla_{\theta} \mathbb{E}_{s,a \sim \rho_{\pi}^{\star}, r, s' \sim P_{R,S}^{\star}(s,a)}\left[\nabla_{\theta} \log p(r, s'|s, a, \theta)\right], \\
&= \nabla_{\theta} \mathbb{E}_{s,a \sim \rho_{\pi}^{\star}, r, s' \sim P_{R,S}^{\star}(s,a)}\left[\frac{\nabla_{\theta} p(r, s'|s, a, \theta)}{p(r, s'|s, a, \theta)}\right], \\
&= \mathbb{E}_{s,a \sim \rho_{\pi}^{\star}, r, s' \sim P_{R,S}^{\star}(s,a)}\left[\nabla_{\theta} \frac{\nabla_{\theta} p(r, s'|s, a, \theta)}{p(r, s'|s, a, \theta)}\right], \\
&= \mathbb{E}_{s,a \sim \rho_{\pi}^{\star}, r, s' \sim P_{R,S}^{\star}(s,a)}\left[\frac{\nabla_{\theta}^2 p(r, s'|s, a, \theta)}{p(r, s'|s, a, \theta)}\right] \\
&\quad - \mathbb{E}_{s,a \sim \rho_{\pi}^{\star}, r, s' \sim P_{R,S}^{\star}(s,a)}\left[\frac{\nabla_{\theta} p(r, s'|s, a, \theta)}{p(r, s'|s, a, \theta)} \frac{\nabla_{\theta} p(r, s'|s, a, \theta)^{\top}}{p(r, s'|s, a, \theta)}\right], \qquad (27)
\end{aligned}$$

Hence at $\theta = \theta_i^{\star}$, the first term of Eq. (27) is:

$$\begin{aligned}
\mathbb{E}_{s,a \sim \rho_{\pi}^{\star}, r, s' \sim P_{R,S}^{\star}(s,a)}\left[\frac{\nabla_{\theta}^2 p(r, s'|s, a, \theta)}{p(r, s'|s, a, \theta_i^{\star})}\right] &= \mathbb{E}_{s,a \sim \rho_{\pi}^{\star}, r, s' \sim P_{R,S}^{\star}(s,a)}\left[\frac{\nabla_{\theta}^2 p(r, s'|s, a, \theta)|_{\theta=\theta_i^{\star}}}{p^{\star}(r, s'|s, a)}\right], \\
&= \mathbb{E}_{s,a \sim \rho_{\pi}^{\star}}\left[\int_{\mathbb{R} \times \mathcal{S}} \nabla_{\theta}^2 p(r, s'|s, a, \theta)|_{\theta=\theta_i^{\star}} d(r, s')\right], \\
&= \mathbb{E}_{s,a \sim \rho_{\pi}^{\star}}\left[\nabla_{\theta}^2 \int_{\mathbb{R} \times \mathcal{S}} p(r, s'|s, a, \theta) d(r, s')|_{\theta=\theta_i^{\star}}\right], \\
&= \nabla_{\theta}^2 1|_{\theta=\theta_i^{\star}}, \\
&= 0,
\end{aligned}$$

hence:

$$\begin{aligned}
H(\theta_i^{\star}) &= 0 - \mathbb{E}_{s,a \sim \rho_{\pi}^{\star}, r, s' \sim P_{R,S}^{\star}(s,a)}\left[\frac{\nabla_{\theta} p(r, s'|s, a, \theta_i^{\star})}{p(r, s'|s, a, \theta_i^{\star})} \frac{\nabla_{\theta} p(r, s'|s, a, \theta_i^{\star})^{\top}}{p(r, s'|s, a, \theta_i^{\star})}\right], \\
&= -\mathbb{E}_{s,a \sim \rho_{\pi}^{\star}, r, s' \sim P_{R,S}^{\star}(s,a)}\left[\nabla_{\theta} \log p(r, s'|s, a, \theta_i^{\star}) \nabla_{\theta} \log p(r, s'|s, a, \theta_i^{\star})^{\top}\right], \\
&= -\Sigma_i^g.
\end{aligned}$$

Using this result, each $\mathrm{Tr}\left(\Sigma_i^g H_i^{\star -1}\right) = \mathrm{Tr}\left(-I\right) = -d$. Substituting y ields:

$$\mathbb{E}_{\mathcal{D}_N \sim P^\star_{\text{Data}}}\left[\text{Regret}(\mathcal{M}^\star, \mathcal{D}_N)\right] \leq 2\mathcal{R}_{\max} \cdot \sqrt{1 - \exp\left(-C\frac{(k+1)d}{(1-\gamma)N}\right)},$$
$$\leq 2\mathcal{R}_{\max} \cdot \sqrt{1 - \exp\left(-C'\frac{d}{(1-\gamma)N}\right)},$$

for some $0 < C' < \infty$ and sufficiently large N, as required. □

We note that Theorem 2 applies to the Gaussian world model introduced in Section 5.2 with neural network mean functions with C^2-continuous activations (tanh, identity, sigmoid, softplus, SiLU, SELU, GELU...) using a Gaussian or uniform prior truncated to a compact parameter space and similarly well-behaved parametric models. The resulting differences in performance only arise from the choice of prior, model representability and coverage of dataset, which affect Bayesian and frequentist methods equally. The information rate coincides with the optimal 'minimax' convergence rate of frequentist parametric density estimators (Yang & Barron, 1999; Bilodeau et al., 2021). Similar results for the information rate have been found for nonparametric models such a Gaussian processes (van der Vaart & van Zanten, 2011), which converge at slower rates.

Using our result in Theorem 2, we plot the normalised regret bound (i.e. taking $\mathcal{R}_{\max} = 0.5$) in Ineq. 20 for increasing dimensionality (blue) and decreasing γ (copper) in Fig. 5. Our bound reveals an S-shaped curve with three distinct phases as number of data points N increases: an initial plateau, a sudden decrease in regret follow by a slow exponential decay towards a regret of zero. The plateau indicates that a minimum amount of data is needed before any benefit can be realised in terms of regret. This is to be expected because initially the only information about the parameter values is given by the prior, which has no guarantee of accuracy under our analysis. Once a threshold of data points has been reached, the data can start to overwhelm the prior, resulting in a sudden decrease in regret. The higher the dimensionality of the model, the greater this data limit is - represented in Fig. 5 by the plateau length increasing with greater d (blue curves). Due to overspecification in models, this limit is likely to be set by the effective dimension of the problem (which may be much lower than d) as many parameters will be redundant, however the effective dimension is typically not possible to ascertain a priori. Finally, we observe that increasing the discount factor γ leads to a longer regret plateau (copper curves) due to any error in the model dynamics being compounded over a longer horizon at test time.

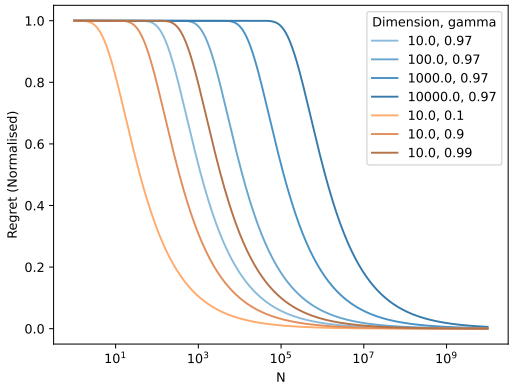


Figure 5: Normalised Regret Curves for $C = 1$

D.4 Extensions for Model Misspecification and Sub-optimal Policy Learning

We now generalise our theorems to include the effects of model misspecification, that is models that cannot fully represent the true environment dynamics, and sub-optimal Bayesian policy learning, that is the effect of using a policy that does not fully optimise the Bayesian RL objective. We use the dagger notation to denote the maximum cross entropy parametrisation:

$$\theta_i^\dagger \in \arg\max_{\theta\in\Theta} \ell(\theta) = \arg\min_{\theta\in\Theta} \mathbb{E}_{s,a\sim\rho^\star_\pi, r, s'\sim P^\star_{R,S}(s,a)}\left[\mathrm{KL}\left(P^\star_{R,S}(s,a) | P_{R,S}(s,a,\theta)\right)\right].$$

To characterise the degree of model misspecification, we use the KL divergence:

$$\epsilon_{\text{miss}} \coloneqq \min_{\theta\in\Theta} \mathbb{E}_{s,a\sim\rho^\star_\pi, r, s'\sim P^\star_{R,S}(s,a)}\left[\mathrm{KL}\left(P^\star_{R,S}(s,a) | P_{R,S}(s,a,\theta)\right)\right].$$

We also introduce the following simplifying dagger notation for the expected cross entropy and corresponding gradients and Hessian under the optimal parameter:

$$\begin{aligned}\ell^\dagger &\coloneqq \max_{\theta\in\Theta} \ell(\theta) = \mathbb{E}_{s,a\sim\rho^\star_\pi, r,s'\sim P^\star_{R,S}(s,a)}\left[\log p(r,s'|s,a,\theta^\dagger_i)\right],\\ g^\dagger_{i,N} &\coloneqq \sqrt{N}\nabla_\theta \ell_N(\theta)\big|_{\theta=\theta^\dagger_i},\\ H^\dagger_i &\coloneqq \nabla^2_\theta \ell(\theta)\big|_{\theta=\theta^\dagger_i}.\end{aligned}$$

We now relax Assumption 2 to allow for model misspecification:
Assumption 2. *We assume that:*

I. The maximum likelihood is finite $\left|\ell^\dagger\right| < \infty$ *and* $\left|\ell^\dagger - \ell(\theta)\right|$ *is bounded* P_Θ*-almost surely.*

II. $\ell_N(\theta)$ *and* $\ell(\theta)$ *are* C^2*-continuous in* θ*.*

III. There are $K < \infty$ *maximising points* $\theta^\dagger_i$*:*

$$\{\theta^\dagger_1, \theta^\dagger_2, \dots \theta^\dagger_K\} = \arg\max_{\theta\in\Theta} \ell(\theta).$$

For each maximiser $\theta^\dagger_i$*, there exists a small region* $\Theta^\dagger_i \coloneqq \{\theta \in \Theta | \|\theta^\dagger_i - \theta\| \leq \epsilon\}$ *for some* $\epsilon > 0$ *such that* $\theta^\dagger_i$ *is the unique maximiser in* $\Theta^\dagger_i$*,* $\theta^\dagger_i$ *is in the interior of* $\theta^\dagger_i$*,* $\nabla^2_\theta \ell(\theta^\dagger_i)$ *is negative definite, invertible and the regions are disjoint:* $\bigcap_{i=1}^K \Theta^\dagger_i = \varnothing$*.*

IV. The prior $p(\theta)$ *is Lipschitz continuous in* θ *with support over* Θ*.*

V. The sampling regime ensures that the strong law of large numbers holds for all maximisers $\theta^\dagger_i$ *for the Hessian, and uniformly for* $\theta \in \Theta$ *for the likelihood, that is:*

$$\ell_N(\theta) \xrightarrow{\text{Unif. a.s.}} \ell(\theta), \quad \nabla^2_\theta \ell_N(\theta^\dagger_i) \xrightarrow{\text{a.s.}} \nabla^2_\theta \ell(\theta^\dagger_i).$$

The central limit theorem applies to the gradient at each $\theta^\dagger_i$*, that is:*

$$\sqrt{N}\nabla_\theta \ell_N(\theta^\dagger_i) \xrightarrow{d} \mathcal{N}(0, \Sigma^g_i),$$

where $\Sigma^g_i = \mathbb{E}_{s,a\sim\rho^\star_\pi, r,s'\sim P^\star_{R,S}(s,a)}\left[\nabla_\theta \log p(r,s'|s,a,\theta^\dagger_i)\nabla_\theta \log p(r,s'|s,a,\theta^\dagger_i)^\top\right]$ *with* $\|\Sigma^g_i\| < \infty$*.*

Finally, we account for let the Bayes sub-optimality be defined as

$$\epsilon_{\text{Bayes}} \coloneqq \left| J^{\pi^\star_{\text{Bayes}}}_{\text{Bayes}}(P_\Phi(\mathcal{D}_N) - J^{\hat{\pi}}_{\text{Bayes}}(P_\Phi(\mathcal{D}_N)\right|. \tag{28}$$

Lemma 5. *Let* $\mathcal{R}_{\max} \coloneqq \frac{(r_{max} - r_{min})}{1-\gamma}$ *denote the maximum possible regret for the MDP and the Bayes sub-optimality* ϵ_{Bayes} *be defined as in Eq. (28). For a prior* $P_\Theta(\mathcal{D}_N)$*, the true regret can be bounded as:*

$$\text{Regret}(\mathcal{M}^\star, \mathcal{D}_N) \leq 2\mathcal{R}_{\max} \cdot \sup_\pi \mathbb{E}_{i\sim\mathcal{G}(\gamma)}\left[TV\left(P^\star_{i+1,\pi} \| P^\pi_{i+1}(\mathcal{D}_N)\right)\right] + \epsilon_{\text{Bayes}}.$$

Proof. We start from the definition of the true regret under Bayes sub-optimality:

$$\begin{aligned}
\text{Regret}(\mathcal{M}^\star, \mathcal{D}_N) &:= J^{\pi^*}(\mathcal{M}^\star) - J^{\hat{\pi}}(\mathcal{M}^\star, \mathcal{D}_N), \\
&= J^{\pi^*}(\mathcal{M}^\star) - J^{\pi^*}_{\text{Bayes}}(P_\Phi(\mathcal{D}_N)) + J^{\pi^*}_{\text{Bayes}}(P_\Phi(\mathcal{D}_N)) - J^{\hat{\pi}}(\mathcal{M}^\star, \mathcal{D}_N), \\
&\leq J^{\pi^*}(\mathcal{M}^\star) - J^{\pi^*}_{\text{Bayes}}(P_\Phi(\mathcal{D}_N)) + J^{\pi^\star_{\text{Bayes}}}_{\text{Bayes}}(P_\Phi(\mathcal{D}_N)) - J^{\hat{\pi}}(\mathcal{M}^\star, \mathcal{D}_N), \\
&= J^{\pi^*}(\mathcal{M}^\star) - J^{\pi^*}_{\text{Bayes}}(P_\Phi(\mathcal{D}_N)) + J^{\hat{\pi}}_{\text{Bayes}}(P_\Phi(\mathcal{D}_N)) - J^{\hat{\pi}}(\mathcal{M}^\star, \mathcal{D}_N) \\
&\quad + J^{\pi^\star_{\text{Bayes}}}_{\text{Bayes}}(P_\Phi(\mathcal{D}_N)) - J^{\hat{\pi}}_{\text{Bayes}}(P_\Phi(\mathcal{D}_N)), \\
&\leq \sup_\pi \left| J^\pi(\mathcal{M}^\star) - J^\pi_{\text{Bayes}}(P_\Theta(\mathcal{D}_N)) \right| + \sup_\pi \left| J^\pi_{\text{Bayes}}(P_\Theta(\mathcal{D}_N)) - J^\pi(\mathcal{M}^\star) \right| \\
&\quad + \left| J^{\pi^\star_{\text{Bayes}}}_{\text{Bayes}}(P_\Phi(\mathcal{D}_N) - J^{\hat{\pi}}_{\text{Bayes}}(P_\Phi(\mathcal{D}_N) \right|, \\
&= 2 \sup_\pi \left| J^\pi(\mathcal{M}^\star) - J^\pi_{\text{Bayes}}(P_\Theta(\mathcal{D}_N)) \right| + \epsilon_{\text{Bayes}},
\end{aligned}$$

where the second line follows from $J^{\pi^*}_{\text{Bayes}}(P_\Phi(\mathcal{D}_N)) \leq J^{\pi^\star_{\text{Bayes}}}_{\text{Bayes}}(P_\Phi(\mathcal{D}_N))$ by definition. We then bound $\left| J^\pi(\mathcal{M}^\star) - J^\pi_{\text{Bayes}}(P_\Theta(\mathcal{D}_N)) \right|$ using Lemma 2 to obtain our desired result. □

Theorem 3. *Let the data be drawn from the underlying true distribution $\mathcal{D}_N \sim P^\star_{Data}$. Under Assumption 2, there exists some constant $0 < C < \infty$ such that for sufficiently large N:*

$$\mathbb{E}_{\mathcal{D}_N \sim P^\star_{Data}}[\text{Regret}(\mathcal{M}^\star, \mathcal{D}_N)] \leq 2\mathcal{R}_{\max} \cdot \exp\left(1 - \sqrt{C\left(\frac{d}{(1-\gamma)N} + \frac{\epsilon_{\text{miss}}}{1-\gamma}\right)}\right) + \epsilon_{\text{Bayes}}.$$

Proof. Starting with Lemma 5, we obtain:

$$\text{Regret}(\mathcal{M}^\star, \mathcal{D}_N) \leq 2\mathcal{R}_{\max} \cdot \sqrt{1 - \exp\left(-C\left(\frac{d^2}{(1-\gamma)N} + \frac{\epsilon_{\text{miss}}}{(1-\gamma)}\right)\right)} + \epsilon_{\text{Bayes}}.$$

Next we apply Theorem 1 to bound the first term, obtaining:

$$\text{Regret}(\mathcal{M}^\star, \mathcal{D}_N) \leq 2\mathcal{R}_{\max} \cdot \sup_\pi \sqrt{1 - \exp\left(-\frac{\mathcal{I}^\pi_N}{1-\gamma}\right)} + \epsilon_{\text{Bayes}}.$$

We write the PIL to include misspecification as:

$$
\begin{aligned}
\mathcal{I}_N^\pi :=& \mathbb{E}_{s,a\sim\rho_\pi^\star}\left[\mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\mathrm{KL}\left(P_{R,S}^\star(s,a)\|P_{R,S}(s,a,\theta)\right)\right]\right], \\
=& \mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\mathbb{E}_{s,a\sim\rho_\pi^\star,r,s'\sim P_{R,S}^\star(s,a)}\left[\log p(r,s'|s,a,\theta^\star) - \log p(r,s'|s,a,\theta)\right]\right], \\
=& \mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\Big[\mathbb{E}_{s,a\sim\rho_\pi^\star,r,s'\sim P_{R,S}^\star(s,a)}[\log p(r,s'|s,a,\theta^\star) - \log p(r,s'|s,a,\theta_i^\dagger) \\
&\quad + \log p(r,s'|s,a,\theta_i^\dagger) - \log p(r,s'|s,a,\theta)]\Big], \\
=& \mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\Big[\mathbb{E}_{s,a\sim\rho_\pi^\star,r,s'\sim P_{R,S}^\star(s,a)}\left[\log p(r,s'|s,a,\theta^\star) - \log p(r,s'|s,a,\theta_i^\dagger)\right] \\
&\quad + \mathbb{E}_{s,a\sim\rho_\pi^\star,r,s'\sim P_{R,S}^\star(s,a)}\left[\log p(r,s'|s,a,\theta_i^\dagger) - \log p(r,s'|s,a,\theta)\right]\Big], \\
=& \mathbb{E}_{s,a\sim\rho_\pi^\star,r,s'\sim P_{R,S}^\star(s,a)}\left[\log p(r,s'|s,a,\theta^\star) - \log p(r,s'|s,a,\theta_i^\dagger)\right] \\
&\quad + \mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\mathbb{E}_{s,a\sim\rho_\pi^\star,r,s'\sim P_{R,S}^\star(s,a)}\left[\log p(r,s'|s,a,\theta_i^\dagger) - \log p(r,s'|s,a,\theta)\right]\right], \\
=& \mathbb{E}_{s,a\sim\rho_\pi^\star,r,s'\sim P_{R,S}^\star(s,a)}\left[\mathrm{KL}\left(P_{R,S}^\star(s,a)|P_{R,S}(s,a,\theta_i^\dagger)\right)\right] \\
&\quad + \mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\mathbb{E}_{s,a\sim\rho_\pi^\star,r,s'\sim P_{R,S}^\star(s,a)}\left[\log p(r,s'|s,a,\theta_i^\dagger) - \log p(r,s'|s,a,\theta)\right]\right], \\
=& \mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\ell^\dagger - \ell(\theta)\right] + \mathbb{E}_{s,a\sim\rho_\pi^\star,r,s'\sim P_{R,S}^\star(s,a)}\left[\mathrm{KL}\left(P_{R,S}^\star(s,a)|P_{R,S}(s,a,\theta_i^\dagger)\right)\right], \\
=& \mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\ell^\dagger - \ell(\theta)\right] + \min_\theta \mathbb{E}_{s,a\sim\rho_\pi^\star,r,s'\sim P_{R,S}^\star(s,a)}\left[\mathrm{KL}\left(P_{R,S}^\star(s,a)|P_{R,S}(s,a,\theta)\right)\right], \\
=& \mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\ell^\dagger - \ell(\theta)\right] + \epsilon_{\text{miss}}.
\end{aligned}
$$

To bound the first term $\mathbb{E}_{\theta\sim P_\Theta(\mathcal{D}_N)}\left[\ell^\dagger - \ell(\theta)\right]$, we follow the remainder of the proof of Theorem 2 to Ineq. 26, replacing $\star$ notion with $\dagger$ to yield:

$$
\mathcal{I}_N^\pi = \epsilon_{\text{miss}} + \mathcal{O}\left(\frac{d - \sum_{i=1}^K {g_{i,N}^\dagger}^\top {H_i^\dagger}^{-1} g_{i,N}^\dagger}{N}\right). \tag{29}
$$

almost surely. As $f(x) := 2\mathcal{R}_{\max}\cdot\sqrt{1-\exp\left(-\frac{x}{(1-\gamma)}\right)}$ is monotonic in x and $\frac{d-\sum_{i=1}^K {g_{i,N}^\dagger}^\top {H_i^\dagger}^{-1} g_{i,N}^\dagger}{N} + \epsilon_{\text{miss}} \geq 0$, Eq. (29) implies there exists some positive $0 < C < \infty$ such that:

$$
\text{Regret}(\mathcal{M}^\dagger, \mathcal{D}_N) \leq 2\mathcal{R}_{\max}\cdot\sqrt{1-\exp\left(-C\left(\frac{d-\sum_{i=1}^K {g_{i,N}^\dagger}^\top {H_i^\dagger}^{-1} g_{i,N}^\dagger}{(1-\gamma)N} + \frac{\epsilon_{\text{miss}}}{(1-\gamma)}\right)\right)},
$$

almost surely for large enough N. Under Assumption 2, $g_{i,N}^\dagger \xrightarrow{d} \mathcal{N}(0,\Sigma_i^g)$. As $f(x)$ is also a bounded, continuous function and concave, we can apply the Portmanteau Theorem (see for example

Bass (2013, Chapter 21.7)) followed by Jensen's inequality to yield:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_N \sim P^{\dagger}_{\text{Data}}}\left[\text{Regret}(\mathcal{M}^{\dagger}, \mathcal{D}_N)\right] \\
&\leq 2\mathcal{R}_{\max} \cdot \mathbb{E}_{g_i \sim \mathcal{N}(0, \Sigma_i^g)}\left[\sqrt{1 - \exp\left(-C\left(\frac{d - \sum_{i=1}^{K} {g_i}^{\top} {H_i^{\dagger}}^{-1} g_i}{(1-\gamma)N} + + \frac{\epsilon_{\text{miss}}}{(1-\gamma)}\right)\right)}\right], \\
\leq & 2\mathcal{R}_{\max} \cdot \sqrt{1 - \exp\left(-C\left(\frac{d - \sum_{i=1}^{K} \mathbb{E}_{g_i \sim \mathcal{N}(0, \Sigma_i^g)}\left[g^{\top} {H_i^{\dagger}}^{-1} g\right]}{(1-\gamma)N} + \frac{\epsilon_{\text{miss}}}{(1-\gamma)}\right)\right)}, \\
= & 2\mathcal{R}_{\max} \cdot \sqrt{1 - \exp\left(-C\left(\frac{d - \sum_{i=1}^{K} \text{Tr}\left(\Sigma_i^g {H_i^{\dagger}}^{-1}\right)}{(1-\gamma)N} + \frac{\epsilon_{\text{miss}}}{(1-\gamma)}\right)\right)}.
\end{aligned}$$

Now $\text{Tr}\left(\Sigma_i^g {H_i^{\star}}^{-1}\right) = \mathcal{O}(d^2)$, hence

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_N \sim P^{\dagger}_{\text{Data}}}\left[\text{Regret}(\mathcal{M}^{\dagger}, \mathcal{D}_N)\right] &\leq 2\mathcal{R}_{\max} \cdot \sqrt{1 - \exp\left(-C\left(\frac{(k+1)d^2}{(1-\gamma)N} + \frac{\epsilon_{\text{miss}}}{(1-\gamma)}\right)\right)}, \\
&\leq 2\mathcal{R}_{\max} \cdot \sqrt{1 - \exp\left(-C'\left(\frac{d^2}{(1-\gamma)N} + \frac{\epsilon_{\text{miss}}}{(1-\gamma)}\right)\right)},
\end{aligned}$$

for some $0 < C' < \infty$ and sufficiently large N, as required. □

Taking the limit $N \to \infty$, we see the residual term due to misspecification and sub-optimal policy learning is:

$$\mathbb{E}_{\mathcal{D}_N \sim P^{\dagger}_{\text{Data}}}\left[\text{Regret}(\mathcal{M}^{\dagger}, \mathcal{D}_N)\right] \leq 2\mathcal{R}_{\max} \cdot \sqrt{1 - \exp\left(-\frac{\epsilon_{\text{miss}} C}{1-\gamma}\right)} + \epsilon_{\text{Bayes}}.$$

We compare against prior work such as MOReL (Kidambi et al., 2020, Corollary 2), where the residual misspecification term is:

$$\frac{4\gamma\mathcal{R}_{\max}}{1-\gamma}\epsilon_{TV}$$

where ϵ_{TV} characterises the misspecification in terms of total variational distance instead of KL divergence of our method. Crucially, we see that our bound is much less sensitive to γ; our bound is $\mathcal{O}(\frac{1}{\sqrt{1-\gamma}})$ in comparison to $\mathcal{O}(\frac{\gamma}{1-\gamma})$ of MOReL meaing our bound is tighter as $\gamma \to 1$.

E Further Results

E.1 SOReL

In practice, each time additional offline data is incorporated, the model, approximate inference and BAMDP hyperparameters should be newly tuned. To avoid too high a computational burden in our experiments, we use fixed model and approximate inference hyperparameters, highlighting in red the region where the approximate regret may be unreliable, and only tune the BAMDP hyperparameters using the approximate regret for one seed and offline dataset size (Fig. 6 in Section E.1).

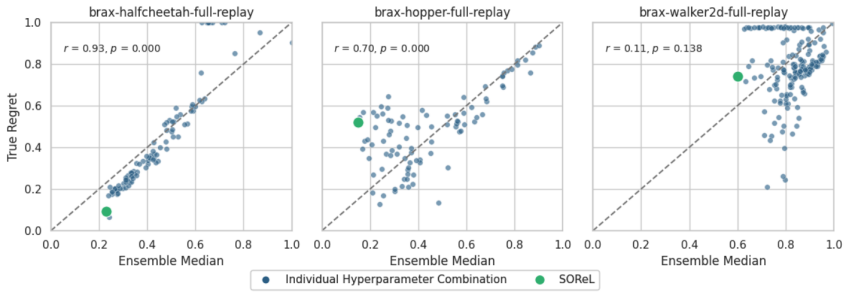


Figure 6: **SOReL BAMDP hyperparameter tuning** for 200,000 randomly sampled transitions of the Brax datasets. SOReL selects the BAMDP hyperparameters that yield the lowest approximate regret (green). Figure corresponds to tuning set ϕ_{III}) in Algorithm 1.

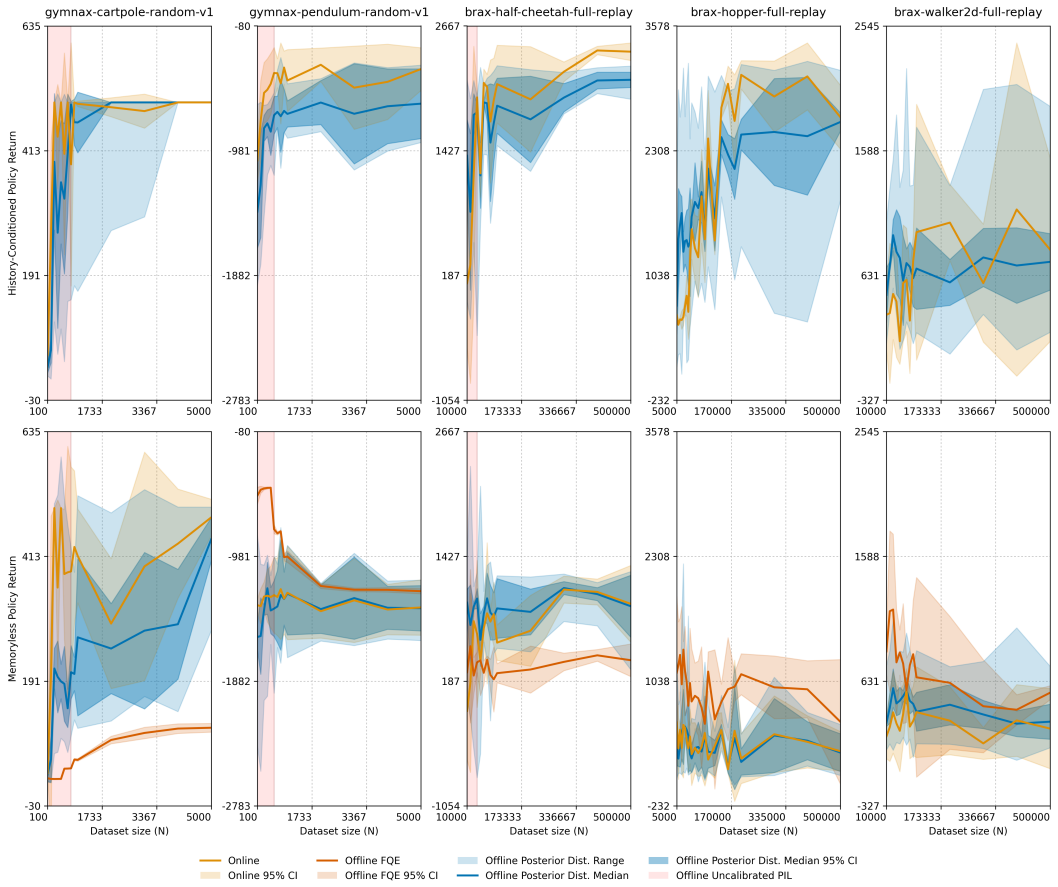


Figure 7: SOReL applied to various tasks (**top**). Approximating the value of a policy for various tasks (**bottom**) (we use the recurrent policy trained using SOReL, but with the history set to 0 across all timesteps (to allow comparison with FQE, since the offline dataset does not contain trajectories)). Light shaded blue shows the **offline approximate return range** across rollouts with environments sampled from the posterior. Dark shaded blue shows the **offline standard deviation of the median rollout** sampled from the posterior. Shaded red indicates where the **PIL is poorly calibrated** $(\mathcal{D}_N, \mathcal{M}^\star) \not\approx \mathcal{V}(\mathcal{D}_N)$ (for a threshold of 0.25), and hence the approximate return may be unreliable. Importantly, a practitioner is aware of this offline. 95% CI over 3 seeds.

E.2 TOReL

Dataset	Alg.	TOReL	FQE	Dual DICE	DR	IS
brax-half-cheetah-full-replay	IQL	0.29 (0.45)				
	ReBRAC	**0.92 (0.00)**				
	MOPO	**0.98 (0.00)**				
	MOReL	**0.93 (0.00)**				
brax-hopper-full-replay	IQL	0.18 (0.635)				
	ReBRAC	**0.98 (0.00)**				
	MOPO	**1.00 (0.00)**				
	MOReL	**0.98 (0.00)**				
brax-walker2d-full-replay	IQL	0.32 (0.41)				
	ReBRAC	**0.81 (0.02)**				
	MOPO	-0.68 (0.13)				
	MOReL	**0.99 (0.00)**				
d4rl-half-cheetah-medium-expert-v2	IQL	0.29 (0.44)	-0.21 (0.59)	-0.47 (0.20)	-0.24 (0.54)	-0.33 (0.39)
	ReBRAC	**0.90 (0.00)**	**0.66 (0.00)**	0.19 (0.17)	0.32 (0.02)	-0.44 (0.01)
	MOPO	0.02 (0.98)				
	MOReL	**0.98 (0.00)**				
d4rl-hopper-medium-v2	IQL	0.29 (0.44)	0.16 (0.67)	0.19 (0.62)	0.26 (0.50)	-0.39 (0.30)
	ReBRAC	**0.98 (0.00)**	0.27 (0.04)	0.34 (0.01)	0.20 (0.16)	0.27 (0.04)
	MOPO	**0.98 (0.00)**				
	MOReL	**0.94 (0.00)**				
d4rl-walker2d-medium-replay-v2	IQL	**0.65 (0.05)**	0.41 (0.28)	0.52 (0.15)	0.19 (0.63)	-0.06 (0.88)
	ReBRAC	**0.78 (0.00)**	**0.89 (0.00)**	-0.04 (0.700)	**0.82 (0.00)**	**0.86 (0.00)**
	MOPO	**1.00 (0.00)**				
	MOReL	-0.53 (0.01)				
d4rl-pen-expert-v1	IQL	0.00 (0.30)	-0.00 (0.40)	0.49 (0.18)	0.51 (0.23)	0.26 (0.50)
	ReBRAC	**0.89 (0.01)**	**0.91 (0.00)**	-0.01 (0.98)	0.26 (0.06)	-0.66 (0.00)
d4rl-hammer-expert-v1	IQL	**0.70 (0.04)**	**0.74 (0.02)**	-0.12 (0.76)	-0.35 (0.35)	-0.40 (0.29)
	ReBRAC	0.30 (0.02)	0.11 (0.42)	-0.14 (0.31)	-0.04 (0.75)	-0.66 (0.00)

Table 1: Pearson correlation r (with p-value in parentheses) between approximate offline and true online returns. **Bold** indicates **strong** ($|r| > 0.50$) and **statistically significant** ($p < 0.05$) correlation.

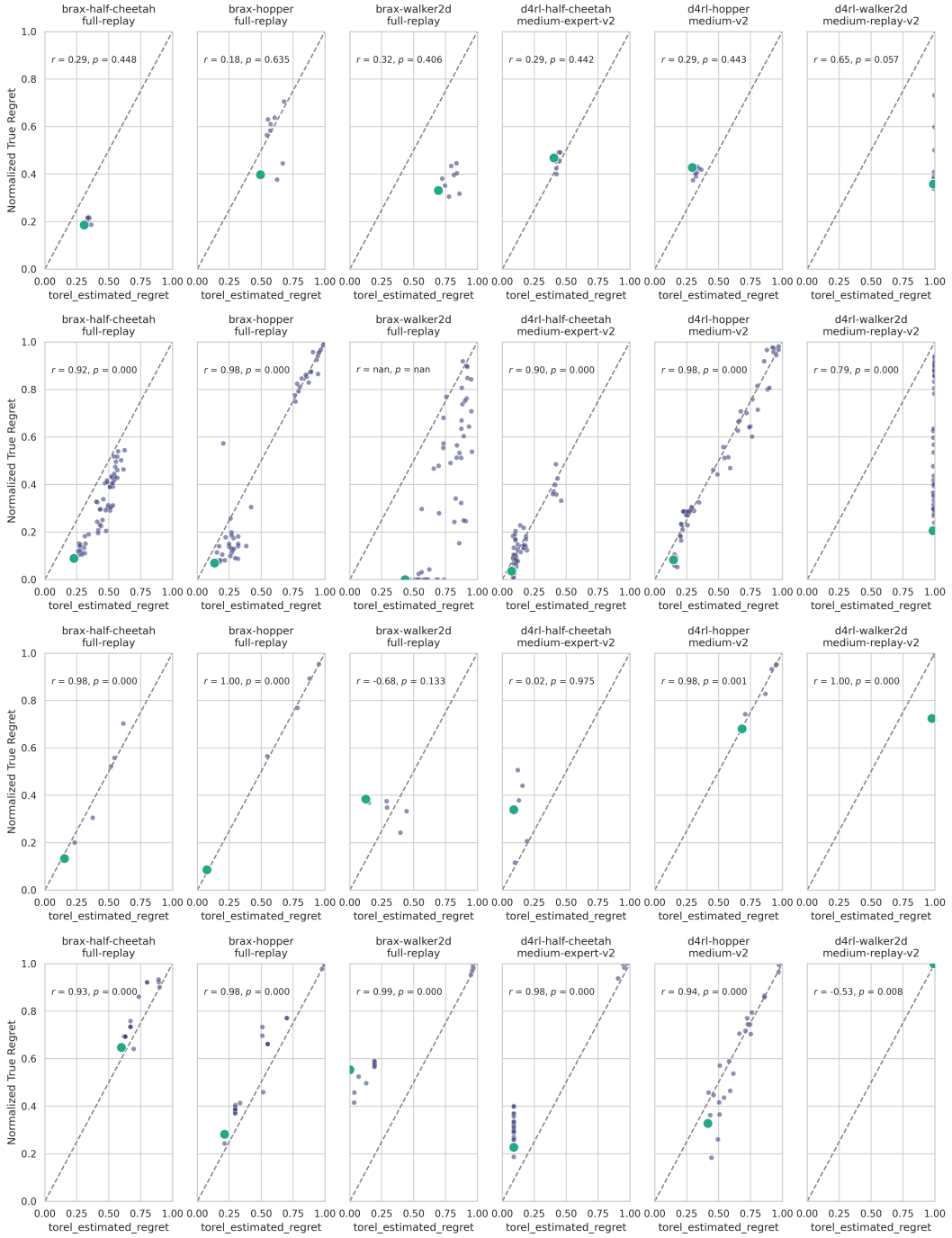
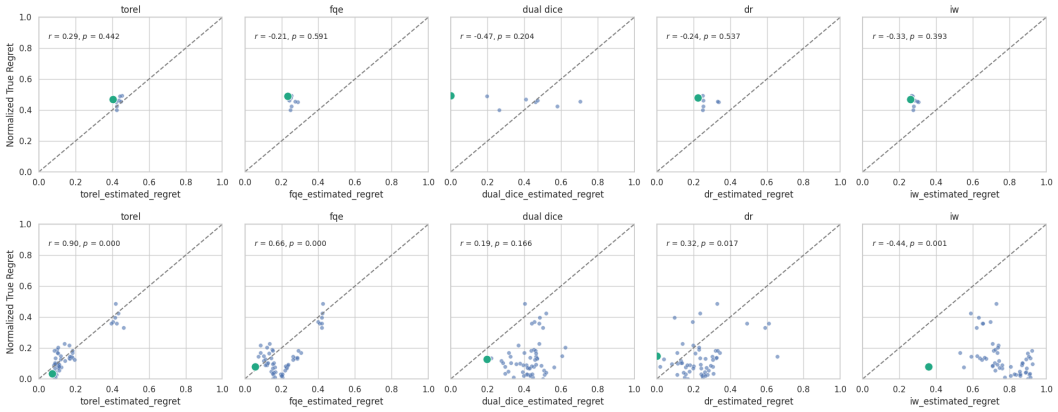
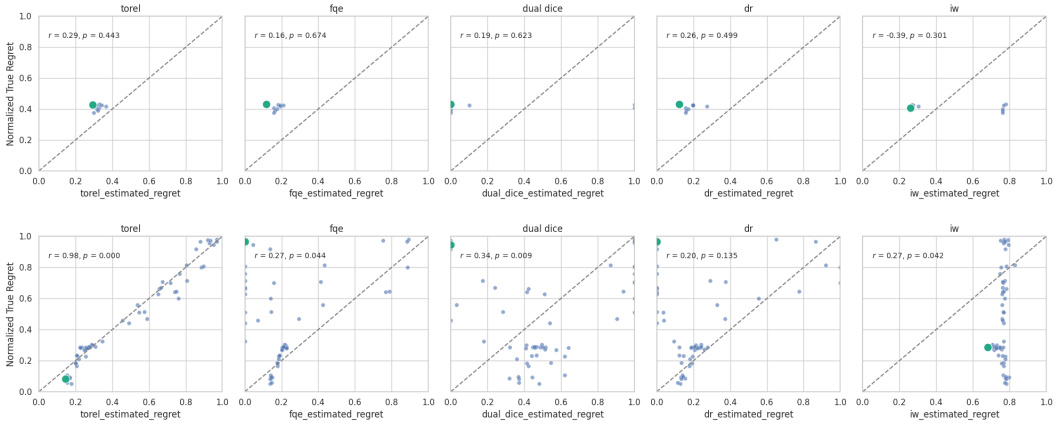


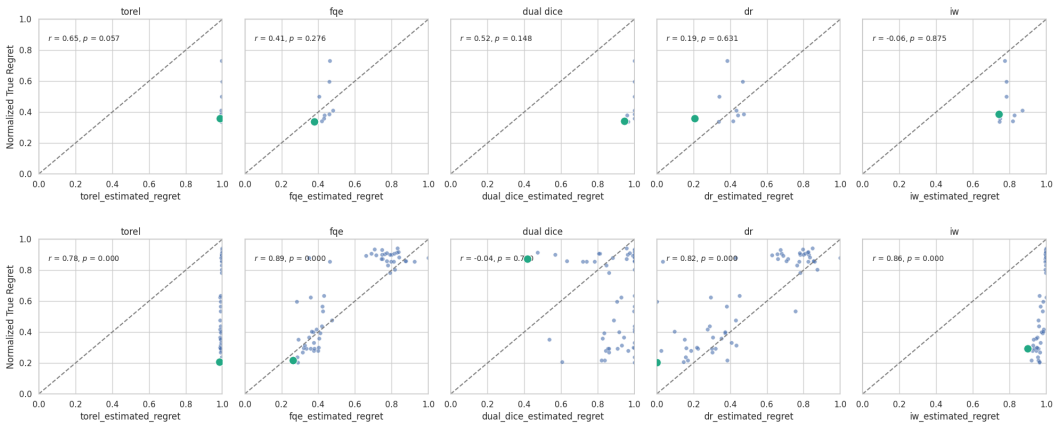
Figure 8: **IQL**, **ReBRAC**, **MOPO** and **MOReL** (**left** to **right**) for brax-half-cheetah-full-replay, brax-hopper-full-replay, brax-walker2d-full-replay, d4rl-half-cheetah-medium-expert-v2, d4rl-hopper-medium-v2, d4rl-walker2d-medium-replay-v2, (**left** to **right**). Scatter plots (1) comparing TOReL's **offline estimated performance** and **online true performance** across datasets and algorithms.



(a) **IQL** (**top**) and **ReBRAC** (**bottom**), d4rl-half-cheetah-medium-expert-v2

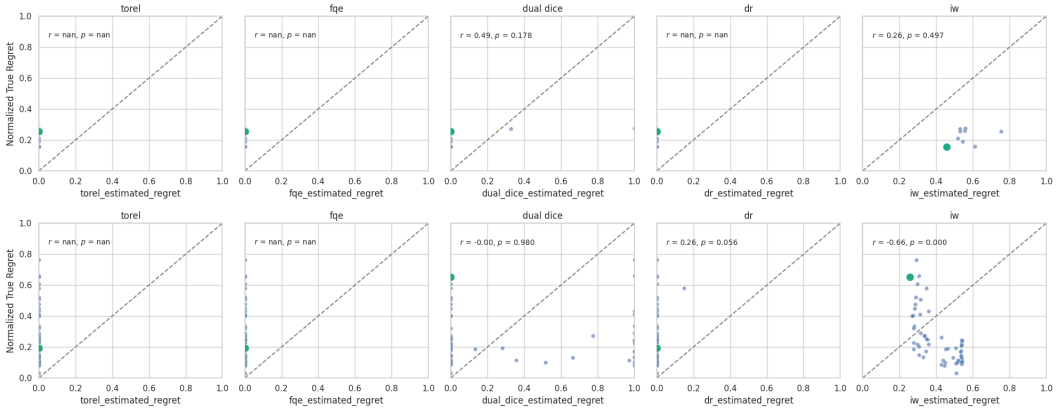


(b) **IQL** (**top**) and **ReBRAC** (**bottom**), d4rl-hopper-medium-v2

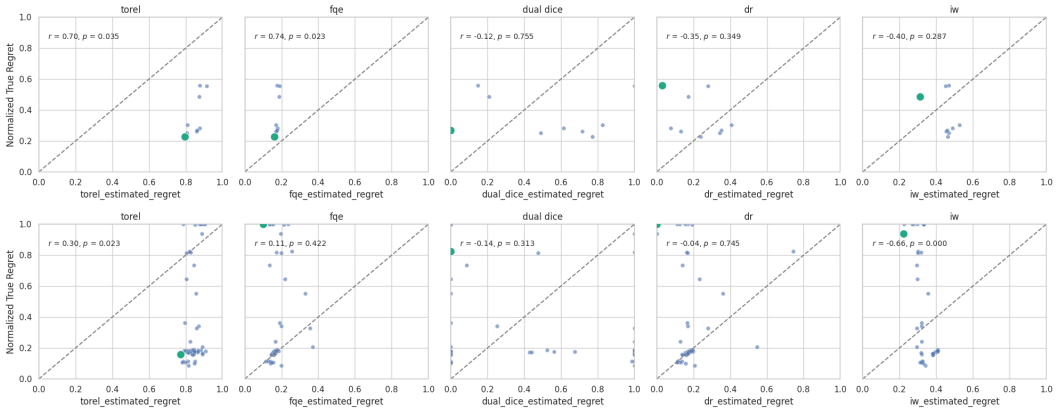


(c) **IQL** (**top**) and **ReBRAC** (**bottom**), d4rl-walker2d-medium-replay-v2

Figure 9: Scatter plots (2) comparing different OPE algorithms' **offline estimated performance** and **online true performance** across datasets and algorithms.



(a) **IQL** (**top**) and **ReBRAC** (**bottom**), d4rl-pen-expert-v1



(b) **IQL** (**top**) and **ReBRAC** (**bottom**), d4rl-hammer-expert-v1

Figure 10: Scatter plots (3) comparing different OPE algorithms' **offline estimated performance** and **online true performance** across datasets and algorithms.

F Implementation Details

Our implementations of SOReL and TOReL are publicly available.

F.1 Diverse full-replay datasets

Without a model prior, the offline dataset must include transitions from poor, medium and expert regions of performance. For the brax environments, we collect our own full-replay datasets to ensure that this is the case. We arbitrarily choose the hyperparameters, simply requiring that the agent spends sufficient time in all three regions of performance. The training curves obtained while collecting the offline datasets are given in Figure 11. The Gymnax environments are simple enough that collecting a dataset using an ensemble of randomly initialised policies leads to sufficient coverage across all three regions of performance.

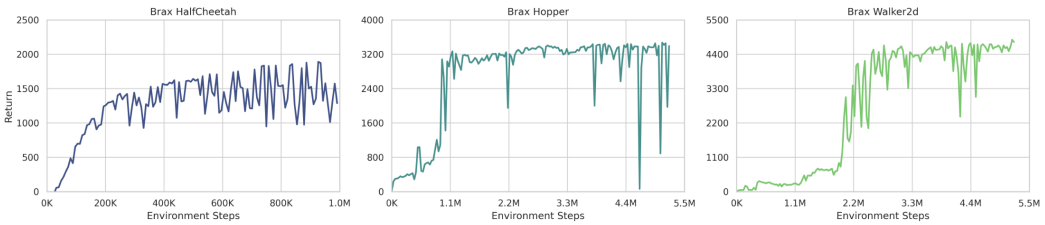


Figure 11: Training curves while **collecting the brax-full-replay** offline datasets. We ensure that the agent spend sufficient time in poor, medium and expert regions of performance such that the offline dataset captures diverse transitions.

F.2 World Model and RP Approximate Inference

To ensure compatibility with Unifloral implementations (Jackson et al., 2025), our world model is a variation of the Gaussian World Model presented in 5.2, but amended to predict the change in state $\Delta := s' - s$ rather than the absolute next state s'.

When rolling out sequences of trajectories on which to train our (Bayes-optimal) policy, we uniformly sample a model from the ensemble of elite models and then sample the transition from the corresponding Gaussian output distribution.

Our ensemble consists of multilayer perceptrons (MLPs) with ReLU activation, which we train using negative log-likelihood loss derived in Section B. Training the models in parallel allows us to simultaneously optimise maximum and minimum (log) variance parameters for each dimension across the model ensemble, which we use to soft-clamp the (log) variances output by the individual models. This prevents any individual model becoming overly confident or too uncertain in one dimension. All models in our ensemble have identical structure, but are initialised differently. The maximum and minimum log-variance terms are initialised at constants. The exact loss function and ensemble dynamics model are the same as the one implemented by Jackson et al. (2025), but we use an Adam optimiser (Kingma & Ba, 2014) with cosine learning rate schedule rather than constant learning rate. A percentage of the available offline dataset is used as a validation set to calculate the PIL. At the end of training, only a subset of elite models are retained, based on their validation MSE. Although the current implementation uses hard-coded reset and termination conditions during model rollouts, the dynamics model could naturally be extended to learn reset and termination heads. When sampling transitions, we conservatively clip the rewards to remain within the support of the offline dataset distribution.

Hyperparameter	Gymnax	Brax/D4RL
Num. layers	3	3
Layer Size	200	200
Activation	ReLU	ReLU
Num. Ensemble Models	7	10
Num. Elite Models	5	8
Log Var. Diff. Coeff.	0.01	0.01
Batch Size	64	256
Num. Epochs	400	400
Learning Rate	0.001	0.001
Learning Rate Schedule	cosine	cosine
Final Learning Rate %	10	10
Weight Decay	2.5e-05	2.5e-05
Validation Split	0.1	0.1

Table 2: World Model Ensemble Dynamics Hyperparameters.

F.3 BAMDP Solver

We use the RNN-PPO implementation of Lu et al. (2022a), which we amend to be compatible with continuous action spaces. We sweep over the hyperparameters given in Table 3.

Hyperparameter	Value / Sweep Values
Learning rate	[0.0001, 0.0003]
Anneal learning rate	True
Number of environments	[4, 64, 128, 256, 512]
Steps per environment	[32, 64, 128]
Total timesteps	Set to 500,000, 1,000,000 or 50,000,000
Update epochs	[2, 4, 8]
Number of minibatches	[2, 4, 8, 16]
Discount factor (γ)	[0.99, 0.995, 0.998]
GAE lambda	[0.8, 0.9, 0.95]
Clip ϵ	[0.2, 0.3]
Entropy coefficient	[0.000, 0.001, 0.010]
Value function coefficient	0.5
Max gradient norm	[0.5, 1.0]
Layer Size	256
Activation function	tanh
RNN size	Set to 64, 128 or 256
Burn-in Percentage	25

Table 3: RNN-PPO hyperparameters swept over.

Hyperparameter	**gymnax-cartpole-random-v1**	**gymnax-pendulum-random-v1**	**brax-halfcheetah-full-replay**	**brax-hopper-full-replay**	**brax-walker2d-full-replay**
Learning rate	0.0003	0.0003	0.0003	0.0003	0.0003
Anneal learning rate	True	True	True	True	True
Number of environments	4	128	512	512	512
Steps per environment	128	64	64	32	64
Total timesteps	500,000	1,000,000	50,000,000	50,000,000	50,000,000
Update epochs	4	8	8	2	4
Number of minibatches	4	16	16	8	8
Gamma	0.99	0.99	0.99	0.998	0.995
GAE lambda	0.95	0.95	0.95	0.8	0.95
Clip ϵ	0.2	0.2	0.2	0.3	0.2
Entropy coefficient	0.01	0.003	0.003	0.001	0.001
Value function coefficient	0.5	0.5	0.5	0.5	0.5
Max gradient norm	0.5	0.5	0.5	1.0	0.5
Layer Size	256	256	256	256	256
Activation function	tanh	tanh	tanh	tanh	tanh
RNN size	64	128	256	256	256
Burn-in Percentage	25	25	25	25	25

Table 4: RNN-PPO hyperparameters for Gymnax and Brax environments. For computational reasons, we sweep over the hyperparameters of each task once, for a fixed dataset size (1000 datapoints for the Gymnax tasks and 200,000 datapoints for Brax tasks) and choose the hyperparameters corresponding to the lowest approximate regret, which we then use to train the policy of all other dataset sizes. Ideally, we would sweep over all hyperparameters for each dataset size.

F.4 ORL Implementations

We use Jackson et al. (2025)'s implementations of the ORL algorithms. We use their default hyperparameters and sweep over their suggested hyperparameters, which we summarise in Table 5.

	Hyperparameter	Value / Sweep Values
Generic Optimisation	Discount factor γ	0.99
	Polyak averaging coefficient	0.005
IQL	Learning rate	0.0003
	Batch size	256
	Beta	[0.5, 3.0, 10.0]
	τ (expectile)	[0.5, 0.7, 0.9]
	Advantage clip	100.0
ReBRAC	Learning rate	0.001
	Batch size	1024
	Critic BC coefficient	[0, 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.1]
	Actor BC coefficient	[0.0005, 0.001, 0.002, 0.003, 0.03, 0.1, 0.3, 1.0]
	Critic layer norm	true
	Actor layer norm	false
	Observation normalization	false
	Noise clip	0.5
	Policy noise	0.2
	Num critic updates per step	2
MOPO	Learning rate	0.0001
	Batch size	256
	Model retain epochs	5
	Number of critics	10
	Rollout batch size	50000
	Rollout interval	1000
	Rollout length	[1, 3, 5]
	Dataset sample ratio	0.05
	Step penalty coefficient	[1.0, 5.0]
MOReL	Learning rate	0.0001
	Batch size	256
	Model retain epochs	5
	Number of critics	10
	Rollout batch size	50000
	Rollout interval	1000
	Rollout length	5
	Dataset sample ratio	0.01
	Threshold coefficient	[0, 5, 10, 15, 20, 25]
	Termination penalty offset	[-30, -50, -100, -200]

Table 5: Hyperparameters and sweep ranges for IQL, ReBRAC, MOPO, and MOReL.

For the sample efficiency experiments in Fig. 4, we use the default UCB bandit-based hyperparameter-tuning algorithm hyperparameters.

F.5 Experiment Compute Resources

All of our experiments were run within a week using four L40S GPUs.