Best Policy Learning From Trajectory Preference Feedback

Akhil Agnihotri **Rahul Jain**
University of Southern California

Deepak Ramachandran
Google DeepMind

Zheng Wen
OpenAI

Abstract

Reinforcement Learning from Human Feedback (RLHF) has emerged as a powerful approach for aligning generative models, but its reliance on learned reward models makes it vulnerable to mis-specification and reward hacking. Preference-based Reinforcement Learning (PbRL) offers a more robust alternative by directly leveraging noisy binary comparisons over trajectories. We study the best policy identification problem in PbRL, motivated by post-training optimization of generative models, for example, during multi-turn interactions. Learning in this setting combines an offline preference dataset—potentially biased or out-of-distribution and collected from a rater of subpar 'competence'—with online pure exploration, making systematic online learning essential. To this end, we propose Posterior Sampling for Preference Learning (PSPL), a novel algorithm inspired by Top-Two Thompson Sampling that maintains posteriors over the reward model and dynamics. We provide the first Bayesian simple regret guarantees for PbRL and introduce an efficient approximation that outperforms existing baselines on simulation and image generation benchmarks.

1 INTRODUCTION

RLHF has recently become a cornerstone of aligning large generative models with human intent, enabling advances in natural language processing and other domains. However, the standard RLHF pipeline depends on learning a reward model from human annotations, which introduces two critical limitations: misspecification of the reward function and susceptibility to reward hacking (Amodei et al., 2016; Novoseller et al., 2020; Tucker et al., 2020). These issues stem from the inherent difficulty of reducing complex human objectives to a scalar reward, even when learned from data (Sadigh et al., 2017; Wirth et al., 2017).

PbRL provides an appealing alternative by relying on comparative rather than absolute feedback, thereby offering a more direct and robust signal of human intent (Christiano et al., 2017; Saha et al., 2023; Metcalf et al., 2024). In many settings, such preferences are expressed over entire trajectories rather than isolated (final) outcomes, which is especially important in human–robot interactions (Wirth and Fürnkranz, 2013; Casper et al., 2023; Dai et al., 2023), experimentation (Novoseller et al., 2020), and recommendation systems (Bengs et al., 2021; Kaufmann et al., 2023). Similarly, for generative systems, trajectory-level feedback captures the multi-turn nature of interactions more faithfully: when users engage with a large language model (LLM), satisfaction often emerges only after a sequence of exchanges (Shani et al., 2024; Zhou et al., 2024).

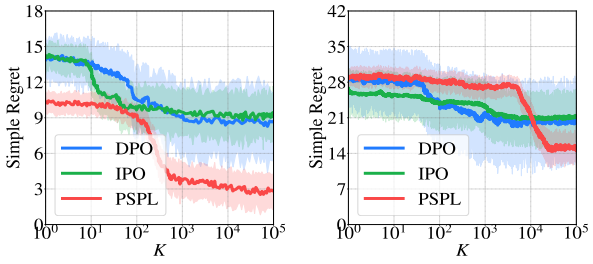
Despite its promise, PbRL in practice often begins with offline preference datasets that are collected off-policy. These datasets are prone to biases and out-of-distribution (OOD) limitations, which can degrade generalization and lead to subpar performance of post-trained models (Zhang and Zanette, 2024; Ming and Li, 2024). This raises a fundamental question: given an offline dataset of trajectory comparisons, how should one systematically supplement it with online preference data to maximize learning efficiency under a constrained budget? This question is particularly important for fine-tuning generative models with human feedback, where large-scale offline data are typically available but careful online exploration can provide higher-value information (Mao et al., 2024; Zhai et al., 2024).

In this work, we propose a systematic framework for leveraging offline datasets to bootstrap online RL algorithms. We demonstrate that, as expected, incorporating offline data consistently improves online learning performance, as reflected in reduced simple regret.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

More interestingly, when the agent is further informed about the 'competence' of the rater generating the offline feedback—equivalently, the behavioral policy underlying the dataset—the resulting informed agent achieves substantially lower simple regret. Finally, we establish that as the rater competence approaches to that of an expert, higher competence levels yield progressively sharper reductions in simple regret compared to baseline methods.

Relevance to RLHF. Our problem setting also connects to LLM alignment, wherein 'best policy identification' (BPI) has emerged as a critical objective. Unlike cumulative regret minimization, which tries to balance exploration and exploitation, BPI ensures that pure exploration leads to the final policy that achieves the highest possible alignment quality (Ouyang et al., 2022; Chung et al., 2024). Moreover, BPI aligns more naturally with human evaluation processes, which often prioritize assessing final system performance over intermediate learning behaviors. This perspective highlights the importance of optimizing simple regret rather than cumulative regret, which aggregates losses over time (Saha et al., 2023; Ye et al., 2024; Zhang et al., 2024a). See Figure 1 for a comparison of PSPL with Direct Preference Optimization (DPO) (Rafailov et al., 2024) and Identity Preference Optimization (IPO) (Gheshlaghi Azar et al., 2024), which shows how on-policy online finetuning helps in BPI.



(a) MountainCar (b) RiverSwim

Figure 1: Comparison of PSPL with current state-of-the-art *offline* finetuning algorithms, DPO and IPO, in two benchmark environments. Online finetuning is necessary for BPI. See Appendix A for more details.

In this paper, we address the problem of BPI for an unknown episodic MDP where both the transition dynamics and reward functions are unknown. We assume that some offline data is available in the form of preferences over pairs of trajectories. And we can then collect additional data by generating pairs of trajectories and seeking contrastive feedback between the two. At the end, we output the best policy we can learn.

Our main contributions are: (i) we introduce a formal framework for best policy learning using offline and online trajectory-preference feedback, (ii) we present a top-two Posterior Sampling algorithm PSPL, the first best policy learning algorithm for this setting and present theoretical bounds on its simple regret, and (iii) we introduce a computationally practical version of PSPL and show that it has excellent empirical performance on benchmark environments as compared to baselines.

Related Work. There has been a fair amount of research for preference-based bandits (contextual and linear) (Singh et al., 2002; Bengs et al., 2021; Busa-Fekete and Hüllermeier, 2014; Saha, 2021; Saha and Gaillard, 2022; Scheid et al., 2024; Cheung and Lyu, 2024). However, even contrary to classical RL (cumulative regret or best policy identification) with numerical reward feedback, only a few works consider incorporating preference feedback in the RL and Bayesian optimization framework (Swamy et al., 2023; Talebi and Maillard, 2018; Zhu et al., 2023; Ye et al., 2024; Song et al., 2022; González et al., 2017). Training RL agents from trajectory-level feedback, available only at the end of each episode, is particularly challenging. Novoseller et al. (2020) and Saha et al. (2023) analyze finite K-episode cumulative regret for this PbRL problem, where two independent trajectories are run, and binary preference feedback is received per episode. While Novoseller et al. (2020) assumes a weaker Gaussian process regression model, Saha et al. (2023) proposes an optimism-based algorithm that is computationally infeasible due to exponential growth in state space variables. Liu et al. (2023) presents a regression-based empirical approach but lacks theoretical guarantees and exploration. Xu et al. (2024); Chen and Tan (2024) address offline PbRL with fixed state-action preference datasets, without considering dataset quality. Ye et al. (2024); Xiong et al. (2023); Li et al. (2024) take a game-theoretic perspective on offline PbRL in RLHF but assume state-action preferences, avoiding credit assignment from trajectory feedback. Hybrid approaches include Agnihotri et al. (2024), which studies PbRL in linear bandits, and Hao et al. (2023), which incorporates numerical rewards in the online phase. *To our knowledge, ours is the first work on BPI in unknown MDPs using trajectory-level preference feedback from a subpar expert.*

2 PRELIMINARIES

Consider a K-episode, H-horizon Markov Decision Process (MDP) setup $\mathcal{M} := (\mathbb{P}_\eta, \mathcal{S}, \mathcal{A}, H, r_\theta, \rho)$, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, $\mathbb{P}_\eta(\cdot \mid s, a)$ are the fixed MDP transition dynamics parameterized by η given a state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, $H \in \mathbb{N}$ is the length of an episode, $r_\theta(\cdot)$ is underlying reward model parameterized by $\theta \in \mathbb{R}^d$, and ρ denotes the initial distribution over states. We let $S := |\mathcal{S}|$ and

$A := |\mathcal{A}|$ to be the cardinalities of the state and action spaces respectively. In addition, we denote the learner by Υ, and let $[N] := \{1, \dots, N\}$ for $N \in \mathbb{N}$.

Under the above MDP setup, a policy $\pi = \{\pi_h\}_{h=1}^H$ is a sequence of mappings from the state space $\mathcal{S}$ to the the probability simplex $\Delta(\mathcal{A})$ over the actions. Specifically, $\pi_h(s, a)$ denotes the probability of selecting action a in state s at step h. In addition, Π is an arbitrary policy class against which performance of the learner will be measured. Finally, we denote a trajectory by concatenation of all states and actions visited during H steps $\tau := (s_1, a_1, \cdots, s_H, a_H)$. At the start of each episode, we assume $s_1 \sim \rho$. For any given (θ, η) pair of reward-transition parameters, the Bellman update equation for a policy $\pi = \{\pi_h(\cdot; \theta, \eta)\}_{h=1}^H$ and for $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ takes the form:

$$Q^\pi_{\theta,\eta,h}(s, a) = r_\theta(s, a) + \mathbb{E}_{s' \sim \mathbb{P}_\eta(\cdot \mid s,a)} \left[V^\pi_{\theta,\eta,h+1}(s') \right], \quad (1)$$

with $V^\pi_{\theta,\eta,h}(s) = \mathbb{E}_{a \sim \pi_h(\cdot|s)}[Q^\pi_{\theta,\eta,h}(s, a)]$ for $h \leqslant H$ and $V^\pi_{\theta,\eta,H+1}(s) = 0 \ \ \forall\ s \in \mathcal{S}$.

Trajectory embedding. For any trajectory τ we assume the existence of a trajectory embedding function $\phi : \Gamma \to \mathbb{R}^d$, where Γ is the set of all possible trajectories of length H and the map ϕ is known to the learner. A special case, which we study in this work, is a decomposed embedding, where $\phi(\tau) = \sum_{h=1}^H \phi(s_h, a_h)$ and $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is a mapping from state-action pairs to $\mathbb{R}^d$. For all trajectories $\tau \in \Gamma$, we assume that $\|\phi(\tau)\|_1 \leqslant B$ for some constant $B > 0$.

Preference Modeling. The learner has access to a rater of arbitrary 'competence'. When presented with two trajectories τ_0 and τ_1, the rater provides feedback in terms of a Bernoulli random variable $Y \in \{0, 1\}$, where $Y = 0$ if the rater prefers τ_0 to τ_1, else $Y = 1$ if τ_1 is preferred to τ_0. Note here that we are working with the setting of sparse feedback which is awarded on a trajectory-level.

We let $\theta \in \mathbb{R}^d$ to be an unknown environment reward vector that the rater has *limited* knowledge of. The rater provides preference feedback based on their 'competence' and their knowledge of the environment reward vector θ. The 'competence' of the rater is characterized by two factors: (i) $\beta \geqslant 0$ is a measure of the *deliberateness* or *surety* of the rater's decision, and (ii) $\lambda > 0$ controls the degree of *knowledgeability* of the rater about θ (see Remark 2.1) . We define $\vartheta \sim N\left(\theta, \mathbf{I}_d/\lambda^2\right)$ ($\mathbf{I}_d$ is a $d \times d$ identity matrix) as the rater's estimate of the true environment vector θ. Pairwise comparison between two trajectories τ_0 and τ_1 from the rater is assumed to follow a Bradley-Terry model (Bradley and Terry, 1952), i.e.,

$$\Pr(Y = 0 \mid \tau_0, \tau_1 \,;\, \vartheta) = \sigma(\beta \langle \phi(\tau_0) - \phi(\tau_1), \vartheta \rangle), \quad (2)$$

where $\sigma : \mathbb{R} \mapsto [0, 1]$ is the logistic link function, i.e., $\sigma(x) = (1 + e^{-x})^{-1}$. This naturally leads to the definition of 'rater score' of a trajectory τ as $g_{\beta,\vartheta}(\tau) := \beta \langle \phi(\tau), \vartheta \rangle$, which is of course dependent on the rater's competence.

Remark 2.1. Intuitively, the parameter $\beta \geqslant 0$ is a measure of the *deliberateness* of the rater's decision: $\beta = 0$ means the rater's decisions are uniformly random, whereas as $\beta \to \infty$ means the rater's decisions are greedy with respect to the trajectory scores. Secondly, λ is the rater's estimate of the true environment reward model based on its knowledgeability i.e., as $\lambda \to \infty$, $\vartheta \to \theta$. In the context of RLHF alignment, λ can be seen as controlling the degree of alignment between a user and the general population from which preferences are aggregated.

Offline Dataset. There is an initial *offline preference* dataset $\mathcal{D}_0$, which is generated by the rater. This offline dataset of size N is a sequence of tuples of the form $\mathcal{D}_0 = \left((\overline{\tau}_n^{(0)}, \overline{\tau}_n^{(1)}, \overline{Y}_n) \right)_{n \in [N]}$, where $\overline{\tau}_n^{(0)}, \overline{\tau}_n^{(1)} \in \Gamma$ are two sampled trajectories, and $\overline{Y}_n \in \{0, 1\}$ indicates the rater's preference.

Learning Objective. Consider an MDP $\mathcal{M}$ with unknown transition model $\mathbb{P}_\eta(s' \mid s, a) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, parameterized by η, and an unknown ground-truth reward function $r_\theta(s, a) : \mathcal{S} \times \mathcal{A} \to [0, 1]$, parameterized by θ, with $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$. One conceivable idea is to assume $r_\theta(\tau)$ as the function $r_\theta(\tau) := \langle \phi(\tau), \theta \rangle$, where $\theta \in \mathbb{R}^d$ is an unknown reward parameter.

The goal of the learner is to identify the optimal policy that maximizes trajectory rewards. Overload notation to denote trajectory rewards by $r_\theta(\tau) := \sum_{h=1}^H r_\theta(s_h, a_h)$. Then denote an optimal policy by $\pi^\star \in \operatorname{argmax}_{\pi \in \Pi} \mathop{\mathbb{E}}_{\tau \sim \pi}[r_\theta(\tau)]$. Consider an optimal policy $\pi^\star$ and any arbitrary policy $\pi \in \Pi$. Then, with $\tau^\star \sim \pi^\star$ and $\tau \sim \pi$, the simple Bayesian regret of π after K online episodes is defined as:

$$\mathcal{SR}^\Upsilon_K(\pi, \pi^\star) := \mathop{\mathbb{E}}_{\tau, \tau^\star}[r_\theta(\tau^\star) - r_\theta(\tau)] \quad (3)$$

In this paper, the objective of the learner Υ then is to design an exploration based online learning algorithm that is *informed* from the offline preference dataset and that which minimizes simple Bayesian regret in Equation (3).

Remark 2.2. One motivation for using pure exploration in the online phase of BPI, rather than active learning approaches that target reward model estimation, is its relevance to RLHF. In practice, online active learning with preference data is infeasible (Chung et al., 2024; Achiam et al., 2023; Anil et al., 2023), since preferences are typically collected in fixed offline batches

and used to train reward models that then guide online learning (e.g., the AutoRater pipeline (Anil et al., 2023)). Thus, pure exploration better reflects this setting, where offline batched preference data is available and only minimal online interaction is possible to learn the best policy.

3 THE PSPL ALGORITHM

For the purpose of pure exploration, the learner Υ is given the opportunity to learn online and generate an *online* dataset of trajectories for which it asks for feedback from the rater. Denote the *online* dataset available to the learner at the beginning of episode $k \in [K]$ as $\mathcal{H}_k = \{(\tau_t^{(0)}, \tau_t^{(1)}, Y_t)\}_{t=1}^{k-1}$. Then the total available dataset for the learner at the beginning of episode k becomes $\mathcal{D}_k = \mathcal{D}_0 \oplus \mathcal{H}_k$, where $\oplus$ denotes concatenation.

Since the parameters θ and η of the reward and transition model are unknown, we assume that the prior distribution over the reward vector θ is a Gaussian distribution $\nu_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$ and over the transition model η is a Dirichlet distribution $\chi_0 \sim \mathrm{Dir}(\boldsymbol{\alpha}_0)$ for each state-action pair, where $\boldsymbol{\alpha}_0$ is a positive real-valued vector of dimension S.

Now, note that the offline dataset *informs* the learner Υ of the MDP transition dynamics and the underlying reward model. This is captured in the *informed* prior (before starting the online phase) over the unknown parameters η and θ. Denote the probability distributions over the transition parameter and reward parameter by $\chi(\eta)$ and $\nu(\theta)$ respectively.

Informed prior for θ. Denoting the *uninformed* prior as $\nu_0(\theta)$, we have the *informed* prior $\nu_1(\theta)$ as,

$$\begin{aligned}\nu_1(\theta) \;&:=\; \nu(\theta \,|\, \mathcal{D}_0) \propto \Pr(\mathcal{D}_0 \,|\, \theta)\nu_0(\theta) \\ &\propto \prod_{n=1}^{N} \Pr(\overline{Y}_n \,|\, \overline{\tau}_n^{(0)}, \overline{\tau}_n^{(1)}, \theta)\; \Pr(\overline{\tau}_n^{(0)} \,|\, \theta)\; \Pr(\overline{\tau}_n^{(1)} \,|\, \theta)\; \nu_0(\theta),\end{aligned} \tag{4}$$

where the second step follows from Equation (2), and that $\overline{\tau}_n^{(0)}$ and $\overline{\tau}_n^{(1)}$ are assumed independent given θ, as is in the context of RLHF, where outputs (trajectories) are conditionally independent given the prompt. It is worth emphasizing that the offline dataset carries information about the reward parameter through the terms $\Pr(\cdot \,|\, \theta)$, which incorporates information about the expert's policy, and thus improves the informativeness of the prior distribution.

Informed prior for η. Denoting the *uninformed* prior as $\chi_0(\eta)$, we have *informed* prior $\chi_1(\eta)$ as,

$$\begin{aligned}\chi_1(\eta) \;&:=\; \chi(\eta \,|\, \mathcal{D}_0) \propto \Pr(\mathcal{D}_0 \,|\, \eta)\, \chi_0(\eta) \\ &\propto \prod_{n=1}^{N} \prod_{j=0}^{1} \prod_{h=1}^{H-1} \mathbb{P}_\eta\left(\overline{s}_{n,h+1}^{(j)} \,|\, \overline{s}_{n,h}^{(j)}, \overline{a}_{n,h}^{(j)}\right)\; \chi_0(\eta),\end{aligned} \tag{5}$$

Algorithm 1 Top-two Posterior Sampling for Preference Learning (PSPL)

Input: Initial dataset $\mathcal{D}_0$, prior on θ as $\nu_0(\theta)$ and η as $\chi_0(\eta)$, horizon H, episodes K.
Construct informed prior $\nu_1(\theta)$ from Equation (4) and $\chi_1(\eta)$ from Equation (5).
for $k = 1, 2, \ldots, K$ **do**
  Sample $\widehat{\eta}_k^{(0)}, \widehat{\eta}_k^{(1)} \sim \chi_k(\eta)$ and $\widehat{\theta}_k^{(0)}, \widehat{\theta}_k^{(1)} \sim \nu_k(\theta)$.
  Compute policies $\pi_k^{(0)}$ using $(\widehat{\eta}_k^{(0)}, \widehat{\theta}_k^{(0)})$ and $\pi_k^{(1)}$ using $(\widehat{\eta}_k^{(1)}, \widehat{\theta}_k^{(1)})$.
  Run two trajectories $\tau_k^{(0)} \sim \pi_k^{(0)}$ and $\tau_k^{(1)} \sim \pi_k^{(1)}$ for H horizon.
  Get feedback Y_k on $\tau_k^{(0)}$ and $\tau_k^{(1)}$, and append to dataset as $\mathcal{D}_k = \mathcal{D}_{k-1} \oplus \left(\tau_k^{(0)}, \tau_k^{(1)}, Y_k\right)$.
  Update posteriors to get $\nu_{k+1}(\theta)$ and $\chi_{k+1}(\eta)$.
end for
Output: Optimal policy $\pi_{K+1}^\star$ computed using MAP estimate from $\chi_{K+1}(\eta)$ and $\nu_{K+1}(\theta)$.

where $\overline{\tau}_n^{(j)} := \{\overline{s}_{n,1}^{(j)}, \overline{a}_{n,1}^{(j)}, \ldots, \overline{s}_{n,H}^{(j)}\}$ for $j \in \{0, 1\}$ is an offline trajectory of length H. Note here that the proportional sign hides the dependence on the offline dataset generating policy (since we do not make any assumptions on this behavourial policy) and the fact that $\overline{Y}_n$ is conditionally independent of dynamics given $\overline{\tau}_n^{(0)}$, $\overline{\tau}_n^{(1)}$.

After constructing informed priors from the offline preference dataset above, the learner begins the online phase for active data collection using pure exploration. The learner maintains posteriors over the true reward and transition kernels, which inherently permit for exploration. In each episode, using samples from these posteriors, the learner computes two policies using value / policy iteration or linear programming, and rolls out two H-horizon trajectories. Posteriors are updated based on the trajectory-level preference feedback, and final policy output is constructed from Maximum-A-Posteriori (MAP) estimate from $\chi_{K+1}(\eta)$ and $\nu_{K+1}(\theta)$, as shown in Algorithm 1.

Remark 3.1. Although there exist optimism-based algorithms which construct confidence sets around the reward and transition kernels, it is known that posterior sampling based algorithms provide superior empirical performance (Ghavamzadeh et al., 2015; Ouyang et al., 2017; Liu et al., 2023). In addition, for our problem setting of best policy identification with noisy trajectory level feedback, posterior sampling is a natural method to incorporate beliefs about the environment.

4 THEORETICAL ANALYSIS OF PSPL

This section focuses on regret analysis of PSPL. The analysis has two main steps: (i) finding a *prior-dependent* upper regret bound in terms of the sub-optimality of any optimal policy estimate $\widehat{\pi}^\star$ constructed from $\mathcal{D}_0$. This part characterizes the online learning phase by upper bounding simple Bayesian regret in terms of the estimate $\widehat{\pi}^\star$ constructed by PSPL before the online phase, and (ii) describing the procedure to construct this $\widehat{\pi}^\star$ based on the attributes of $\mathcal{D}_0$, such as size N and rater competence (λ, β). Proofs of results, if not given, are provided in Appendix A.

4.1 General prior-dependent regret bound

It is natural to expect some regret reduction if an offline preference dataset is available to warm-start the online learning. However, the degree of improvement must depend on the 'quality' of this dataset, for example through its size N or rater competence (λ, β). Thus, analysis involves obtaining a prior-dependent regret bound, which we obtain next.

Lemma 4.1. *For any confidence $\delta_1 \in (0, \frac{1}{3})$, let $\delta_2 \in (c, 1)$ with $c \in (0, 1)$, be the probability that any optimal policy estimate $\widehat{\pi}^\star$ constructed from the offline preference dataset $\mathcal{D}_0$ is ε-optimal with probability at least $(1-\delta_2)$ i.e., $\Pr\left(\mathop{\mathbb{E}}_{s\sim\rho}\left[V^{\pi^\star}_{\theta,\eta,0}(s) - V^{\widehat{\pi}^\star}_{\theta,\eta,0}(s)\right] > \varepsilon\right) < \delta_2$. Then, the simple Bayesian regret of the learner Υ is upper bounded with probability at least $1-3\delta_1$ by,*

$$\mathcal{SR}^{\Upsilon}_K(\pi^\star_{K+1}, \pi^\star) \leqslant \sqrt{\frac{10\delta_2 S^2AH^3 \ln\left(\frac{2KSA}{\delta_1}\right) + 3SAH^2\varepsilon^2}{2K\left(1 + \ln\frac{SAH}{\delta_1}\right) - \ln\frac{SAH}{\delta_1}}} \tag{6}$$

Please see Appendix A.1 for proof. This lemma tells us that if the learner can construct an ε-optimal policy estimate $\widehat{\pi}^\star$ with a probability of $(1-\delta_2)$, then the learner's simple regret decreases as δ_2 decreases. Next, we describe how to incorporate information from $\mathcal{D}_0$ before the online phase.

4.2 Incorporating offline preferences for regret analysis

Before we begin to construct $\widehat{\pi}^\star$, we define the state visitation probability $p^\pi_h(s)$ for any given policy π. Construction of $\widehat{\pi}^\star$ will revolve around classification of states in a planning step based on $p^\pi_h(s)$, which will lead to a reward free exploration strategy to enable learning the optimal policy.

Definition 4.2 (State Visitation Probability)**.** Given $(h, s) \in [H] \times \mathcal{S}$, the state (occupancy measure) and state-action visitation probabilities of a policy π is defined for all $h' \in [h-1]$ as follows:

$$\begin{aligned} p^\pi_h(s) &= \Pr\left(s_h = s \,|\, s_1 \sim \rho, a_{h'} \sim \pi(s_{h'})\right) \\ p^\pi_h(s,a) &= \Pr\left(s_h = s, a_h = a \,|\, s_1 \sim \rho, a_{h'} \sim \pi(s_{h'})\right). \end{aligned} \tag{7}$$

Let $p_{\min} = \min_{h,s} \max_\pi p^\pi_h(s)$, and we assume it is positive. Further, define the infimum probability of any reachable state under $\pi^\star$ as $p^\star_{\min} := \min_{h,s} p^{\pi^\star}_h(s)$ and we assume it positive as well.

We now describe a procedure to construct an estimator $\widehat{\pi}^\star$ of the optimal policy from $\mathcal{D}_0$ such that it is ε-optimal with probability δ_2. For each (θ, η), define a deterministic Markov policy $\pi^\star(\theta,\eta) = \{\pi^\star_h(\cdot;\theta,\eta)\}_{h=1}^H$,

$$\pi^\star_h(s;\theta,\eta) = \begin{cases} \arg\max_a Q_{\theta,\eta,h}(s,a), & \text{if } p^{\pi^\star}_h(s) > 0 \\ \mathring{a} \sim \mathrm{Unif}(\mathcal{A}), & \text{if } p^{\pi^\star}_h(s) = 0, \end{cases} \tag{8}$$

where $Q_{\theta,\eta,h}(\cdot,\cdot)$ is the Q-value function in a MDP with reward-transition parameters (θ,η), and the tiebreaker for the argmax operation is based on any fixed order on actions. It is clear from construction that $\{\pi^\star_h(\cdot;\theta,\eta)\}_{h=1}^H$ is an optimal policy for the MDP with parameters (θ,η). Furthermore, for those states that are impossible to be visited, we choose to take an action uniformly sampled from $\mathcal{A}$.

Construction of $\widehat{\pi}^\star$. To attribute preference of one trajectory to another, we build 'winning' (U^W_h), and 'undecided' (U^U_h) action (sub)sets for each state in the state space for each time step $h \in [H]$. To achieve this, we define the net count of each state-action pair from the offline dataset $\mathcal{D}_0$. Recall that any offline trajectory $\overline{\tau}^{(\cdot)}_n$ is composed of $(\overline{s}^{(\cdot)}_{n,1}, \overline{a}^{(\cdot)}_{n,1}, \ldots, \overline{s}^{(\cdot)}_{n,H})$. Then if the 'winning' counts are defined as $w_h(s,a) = \sum_{n=1}^N \mathbf{I}\{\overline{s}^{(\overline{Y_n})}_{n,h} = s, \overline{a}^{(\overline{Y_n})}_{n,h} = a\}$, and 'losing' counts are defined as $l_h(s,a) = \sum_{n=1}^N \mathbf{I}\{\overline{s}^{(1-\overline{Y_n})}_{n,h} = s, \overline{a}^{(1-\overline{Y_n})}_{n,h} = a\}$, then the 'net' counts are defined as $c_h(s,a) = w_h(s,a) - l_h(s,a)$. The action sets are then constructed as,

$$U^W_h(s) = \{a; c_h(s,a) > 0\}, U^U_h(s) = \{a; c_h(s,a) \leqslant 0\}$$

By construction, if any state-action pair does not appear in $\mathcal{D}_0$, it is attributed to the undecided set U^U_h. Finally, for some fixed $\delta \in (0,1)$, construct the policy estimate $\widehat{\pi}^\star := \{\widehat{\pi}^\star_h\}_{h=1}^H$ as

$$\widehat{\pi}^\star_h(s) := \begin{cases} \mathrm{argmax}_{a \in U^W_h(s)} c_h(s,a) & \text{if } \sum_{a'} c_h(s,a') \geqslant \delta N \\ \mathring{a} \sim \mathrm{Unif}(U^U_h) & \text{if } \sum_{a'} c_h(s,a') < \delta N \end{cases}$$

To ensure that for any state-time pair (s,h) with $p^{\pi^\star}_h(s) > 0$, the optimal action $\pi^\star_h(s;\cdot) \in U^W_h(s)$ with high probability, we first need an upper bound on the 'error' probability of the rater w.r.t any trajectory pair in the offline dataset $\mathcal{D}_0$. Here, the 'error' probability refers to the probability of the event in which the

rater prefers the sub-optimal trajectory (w.r.t. the trajectory score $g(\cdot)$). Hence, for $n \in [N]$, define an event $\mathcal{E}_n := \left\{ \overline{Y}_n \neq \text{argmax}_{i\in\{0,1\}} g_{\beta,\vartheta}\left(\overline{\tau}_n^{(i)}\right) \right\}$, i.e., at the n-th index of the offline preference dataset, the rater preferred the suboptimal trajectory. Then,

Lemma 4.3. *Given a rater with competence (λ, β) such that $\beta > \frac{2\ln\left(2d^{1/2}\right)}{|B\lambda^2 - 2\Delta_{\min}|}$ with $\Delta_{\min} = \min_{n\in[N]} \left| r_\theta(\overline{\tau}_n^{(0)}) - r_\theta(\overline{\tau}_n^{(1)}) \right|$, and an offline preference dataset $\mathcal{D}_0$ of size $N > 2$, we have*

$$\Pr(\mathcal{E}_n \,|\, \beta, \vartheta) \leqslant \exp\left(-\beta B \sqrt{2\ln(2d^{1/2}N)/\lambda} - \beta\Delta_{\min} \right) + \frac{1}{N} \;:=\; \gamma_{\beta,\lambda,N}$$

See Appendix A.1 for proof. This lemma establishes the relationship between the competence of a rater (in terms of (λ, β)) and the probability with which it is sub-optimal in its preference. Observe that as the rater tends to an expert i.e., $\lambda, \beta \to \infty$ and we have increasing offline dataset size N, we get $\Pr(\mathcal{E}_n \,|\, \beta, \vartheta) \to 0$ i.e., the rater *always* prefers the trajectory with higher rewards in an episode, which is intuitive to understand. Using this Lemma 4.3, we can upper bound the simple Bayesian regret of PSPL as below.

Theorem 4.4. *For any confidence $\delta_1 \in (0, \frac{1}{3})$ and offline preference dataset size $N > 2$, the simple Bayesian regret of the learner Υ is upper bounded with probability of at least $1 - 3\delta_1$ by,*

$$\mathcal{SR}_K^{\Upsilon}(\pi^\star_{K+1}, \pi^\star) \leqslant \sqrt{\frac{20\delta_2 S^2 AH^3 \ln\left(\frac{2KSA}{\delta_1}\right)}{2K\left(1 + \ln\frac{SAH}{\delta_1}\right) - \ln\frac{SAH}{\delta_1}}}$$

$$, \textit{with } \delta_2 = 2\exp\left(-N\left(1+\gamma_{\beta,\lambda,N}\right)^2\right) + \exp\left(-\frac{N}{4}(1-\gamma_{\beta,\lambda,N})^3\right).$$

See Appendix A.1 for the proof. For a fixed $N > 2$, and large S and A, the simple regret bound is $\widetilde{\mathcal{O}}\left(\sqrt{S^2AH^3K^{-1}}\right)$. Note that this bound converges to zero exponentially fast as $N \to \infty$ and as the rater tends to an expert (large β, λ). In addition, as the number of episodes K gets large, PSPL is able to identify the best policy with probability at least $(1 - 3\delta_1)$.

Remark 4.5. (i) This paper is different from existing hybrid RL works which usually consider some notion of 'coverability' in the offline dataset $\mathcal{D}_0$ (Wagenmaker and Jamieson, 2022; Song et al., 2022). We instead deal directly with the estimation of the optimal policy, which depends on the visits to each state-action pair in $\mathcal{D}_0$. Rather than dealing with concentratability coefficients, we deal with counts of visits, which provides a simple yet effective way to incorporate coverability of the offline phase into the online phase. If the coverability (counts of visits) is low in $\mathcal{D}_0$, the failure event in Lemma 4.3 has a higher upper bound, which directly impacts the simple regret in Theorem 4.4 negatively. (ii) Since bandit models are a special case of our setting, results specific for them are presented in Appendix A.2.

5 A PRACTICAL APPROXIMATION

The PSPL algorithm introduced above assumes that posterior updates can be solved in closed-form. In practice, the posterior update in equations (4) and (5) is challenging due to loss of conjugacy. Hence, we propose a novel approach using Bayesian bootstrapping to obtain approximate posterior samples. The idea is to perturb the loss function for the maximum a posteriori (MAP) estimate and treat the resulting point estimate as a proxy for an exact posterior sample.

We start with the MAP estimate problem for $(\theta, \vartheta, \eta)$ given the offline and online dataset $\mathcal{D}_k$ at the beginning of episode time k. We show that this is equivalent to minimizing a particular surrogate loss function described below. For any parity of trajectory $j \in \{0, 1\}$ during an episode k, we denote the estimated transition from state-action $s_{k,h}^{(j)}, a_{k,h}^{(j)}$ to state $s_{k,h+1}^{(j)}$ by $\Pr_\eta\left(s_{t,h+1}^{(j)} \,|\, s_{t,h}^{(j)}, a_{t,h}^{(j)}\right)$, where $\Pr_\eta(\cdot)$ is updated based on counts of visits. Note here that as transition dynamics are independent of trajectory parity, we update these counts based on all trajectories in $\mathcal{D}_k$. See Appendix A.3 for proof.

Lemma 5.1. *At episode k, the MAP estimate of $(\theta, \vartheta, \eta)$ can be constructed by solving the following equivalent optimization problem:*

$$\begin{aligned}
(\theta_{opt}, \vartheta_{opt}, \eta_{opt}) &= \underset{\theta,\vartheta,\eta}{\text{argmax}}\ \Pr(\theta, \vartheta, \eta \,|\, \mathcal{D}_k) \\
&\equiv \underset{\theta,\vartheta,\eta}{\text{argmin}}\ \mathcal{L}_1(\theta,\vartheta,\eta) + \mathcal{L}_2(\theta,\vartheta,\eta) + \mathcal{L}_3(\theta,\vartheta,\eta)\ , \textit{where,} \\
\mathcal{L}_1(\theta,\vartheta,\eta) &:= -\sum_{t=1}^{k-1}\left[\beta\langle\tau_t^{(Y_t)}, \vartheta\rangle - \ln\left(e^{\beta\langle\tau_t^{(0)},\vartheta\rangle} + e^{\beta\langle\tau_t^{(1)},\vartheta\rangle}\right)\right. \\
&\qquad \left. + \sum_{j=0}^{1}\sum_{h=1}^{H-1} \ln \Pr_\eta\left(s_{t,h+1}^{(j)} \,|\, s_{t,h}^{(j)}, a_{t,h}^{(j)}\right)\right], \\
\mathcal{L}_2(\theta,\vartheta,\eta) &:= -\sum_{n=1}^{N}\left[\beta\langle\overline{\tau}_n^{(\overline{Y}_n)}, \vartheta\rangle - \ln\left(e^{\beta\langle\overline{\tau}_n^{(0)},\vartheta\rangle} + e^{\beta\langle\overline{\tau}_n^{(1)},\vartheta\rangle}\right)\right], \\
\mathcal{L}_3(\theta,\vartheta,\eta) &:= \frac{\lambda^2}{2}\|\theta - \vartheta\|_2^2 - SA\sum_{i=1}^{S}(\boldsymbol{\alpha}_{0,i} - 1)\ln\eta_i + \frac{1}{2}(\theta - \mu_0)^T\Sigma_0^{-1}(\theta - \mu_0).
\end{aligned}$$

Perturbation of MAP estimate. As mentioned above, the idea now is to *perturb* the loss function in Equation (9) with some noise, so that the MAP point estimates we get from this perturbed surrogate loss function serve as *samples* from a distribution that approximates the true posterior (Osband et al., 2019; Lu and Van Roy, 2017; Qin et al., 2022; Dwaracherla et al., 2022). To that end, we use a perturbation of the 'online' loss function $\mathcal{L}_1(\cdot)$ and of the 'offline' loss function $\mathcal{L}_2(\cdot)$ by multiplicative random weights, and of the 'prior' loss function $\mathcal{L}_3(\cdot)$ by random samples

Algorithm 2 Bootstrapped Posterior Sampling for Preference Learning

Input: Initial dataset $\mathcal{D}_0$, priors $\nu_0(\theta)$ and $\chi_0(\eta)$.
for $k = 1, 2, \ldots, K+1$ **do**
  for $i = 0, 1$ **do**
    Sample a set of perturbations $\mathcal{P}_k^{(i)} = \{(\zeta_t^{(i)}, \omega_n^{(i)}, \theta'^{(i)}, \vartheta'^{(i)})\}$.
    Solve Equation (9) using $\mathcal{P}_k^{(i)}$ to find $(\widehat{\theta}_k^{(i)}, \widehat{\vartheta}_k^{(i)}, \widehat{\eta}_k^{(i)})$.
  end for
  Compute optimal policies $\pi_k^{(0)}$ using $(\widehat{\eta}_k^{(0)}, \widehat{\theta}_k^{(0)})$ and $\pi_k^{(1)}$ using $(\widehat{\eta}_k^{(1)}, \widehat{\theta}_k^{(1)})$.
  Run two trajectories $\tau_k^{(0)} \sim \pi_k^{(0)}$ and $\tau_k^{(1)} \sim \pi_k^{(1)}$ for H horizon.
  Get feedback Y_k on $\tau_k^{(0)}$ and $\tau_k^{(1)}$, and append to dataset as $\mathcal{D}_k = \mathcal{D}_{k-1} \oplus \left(\tau_k^{(0)}, \tau_k^{(1)}, Y_k\right)$.
end for
Output: Optimal policy $\pi_{K+1}^{\star} \equiv \pi_{K+1}^{(0)}$.

from the prior distribution [1]:

(i) *Online perturbation.* Let $\zeta_t \sim \text{Bern}(0.75)$, all i.i.d. Then, the perturbed $\mathcal{L}_1(\cdot)$ becomes,

$$\mathcal{L}_1'(\theta, \vartheta, \eta) = -\sum_{t=1}^{k-1} \zeta_t \left[\beta \langle \tau_t^{(Y_t)}, \vartheta \rangle - \ln \left(e^{\beta \langle \tau_t^{(0)}, \vartheta \rangle} + e^{\beta \langle \tau_t^{(1)}, \vartheta \rangle} \right) + \sum_{j=0}^{1} \sum_{h=1}^{H-1} \ln P_\eta \left(s_{t,h+1}^{(j)} \mid s_{t,h}^{(j)} a_{t,h}^{(j)} \right) \right]$$

(ii) *Offline perturbation.* Let $\omega_n \sim \text{Bern}(0.6)$, all i.i.d. Then, the perturbed $\mathcal{L}_2(\cdot)$ becomes,

$$\mathcal{L}_2'(\theta, \vartheta, \eta) = -\sum_{n=1}^{N} \omega_n \left[\beta \langle \overline{\tau}_n^{(\overline{Y}_n)}, \vartheta \rangle - \ln \left(e^{\beta \langle \overline{\tau}_n^{(0)}, \vartheta \rangle} + e^{\beta \langle \overline{\tau}_n^{(1)}, \vartheta \rangle} \right) \right]$$

(iii) *Prior perturbation.* Let $\theta' \sim \mathcal{N}(\mu_0, \Sigma_0)$, $\vartheta' \sim \mathcal{N}(\mu_0, \mathbf{I}_d/\lambda^2)$, all i.i.d. The perturbed $\mathcal{L}_3(\cdot)$ is,

$$\mathcal{L}_3'(\theta, \vartheta, \eta) = \frac{\lambda^2}{2} \|\theta - \vartheta + \vartheta'\|_2^2 - SA \sum_{i=1}^{S} (\boldsymbol{\alpha}_{0,i} - 1) \ln \eta_i + \frac{1}{2} (\theta - \mu_0 - \theta')^T \Sigma_0^{-1} (\theta - \mu_0 - \theta')$$

Then, for the k^{th} episode, we get the following MAP point estimate from the perturbed loss function,

$$(\widehat{\theta}_k, \widehat{\vartheta}_k, \widehat{\eta}_k) = \underset{\theta, \vartheta, \eta}{\text{argmin}}\ \mathcal{L}_1'(\theta, \vartheta, \eta) + \mathcal{L}_2'(\theta, \vartheta, \eta) + \mathcal{L}_3'(\theta, \vartheta, \eta), \quad (9)$$

which are understood to have a distribution that approximates the actual posterior distribution. Note that this loss function can be extended easily to the setting where the offline dataset comes from *multiple* raters with different (λ_i, β_i) competencies. Specifically, for M raters, there will be M similar terms for $\mathcal{L}_1'(\cdot)$, $\mathcal{L}_2'(\cdot)$, and $\mathcal{L}_3'(\cdot)$. The final algorithm is given as Algorithm 2.

[1] Bern$(\cdot)$ parameters that maximize performance are provided.

6 EMPIRICAL RESULTS

We now present results on the empirical performance of Bootstrapped PSPL algorithm. We first demonstrate the effectiveness of PSPL on synthetic simulation benchmarks and then on realworld datasets. We study: (i) How much is the reduction in simple Bayesian regret after warm starting with an offline dataset? (ii) How much does the competence (λ and β) of the rater who generated the offline preferences affect simple regret? (iii) Is PSPL robust to mis-specification of λ and β?

Baselines. To evaluate the Bootstrapped PSPL algorithm, we consider the following baselines: (i) Logistic Preference based Reinforcement Learning (LPbRL) (Saha et al., 2023), and (ii) Dueling Posterior Sampling (DPS) (Novoseller et al., 2020). LPbRL does not specify how to incorporate prior offline data in to the optimization problem, and hence, no data has been used to warm start LPbRL, but, we initialize the transition and reward models in DPS using $\mathcal{D}_0$. We run and validate the performance of all algorithms in the RiverSwim (Strehl and Littman, 2008) and MountainCar (Moore, 1990) environments. See Figure 3 for empirical Bayesian regret comparison.

Value of Offline Preferences. We first understand the impact of $\mathcal{D}_0$ on the performance of PSPL as the parameters (β, λ and N) vary. Figure 2 shows that as λ increases (the rater has a better estimate of the reward model θ), the regret reduces substantially. Also notice that (for fixed λ and N) as β increases, the regret reduces substantially. Lastly, for fixed β and λ, as dataset size N increases, even with a 'mediocre' expert ($\beta = 5$) the regret reduces substantially. Overall, notice that PSPL correctly incorporates rater competence, which is intuitive to understand since in practice, rater feedback is imperfect, so weighting by competence is crucial.

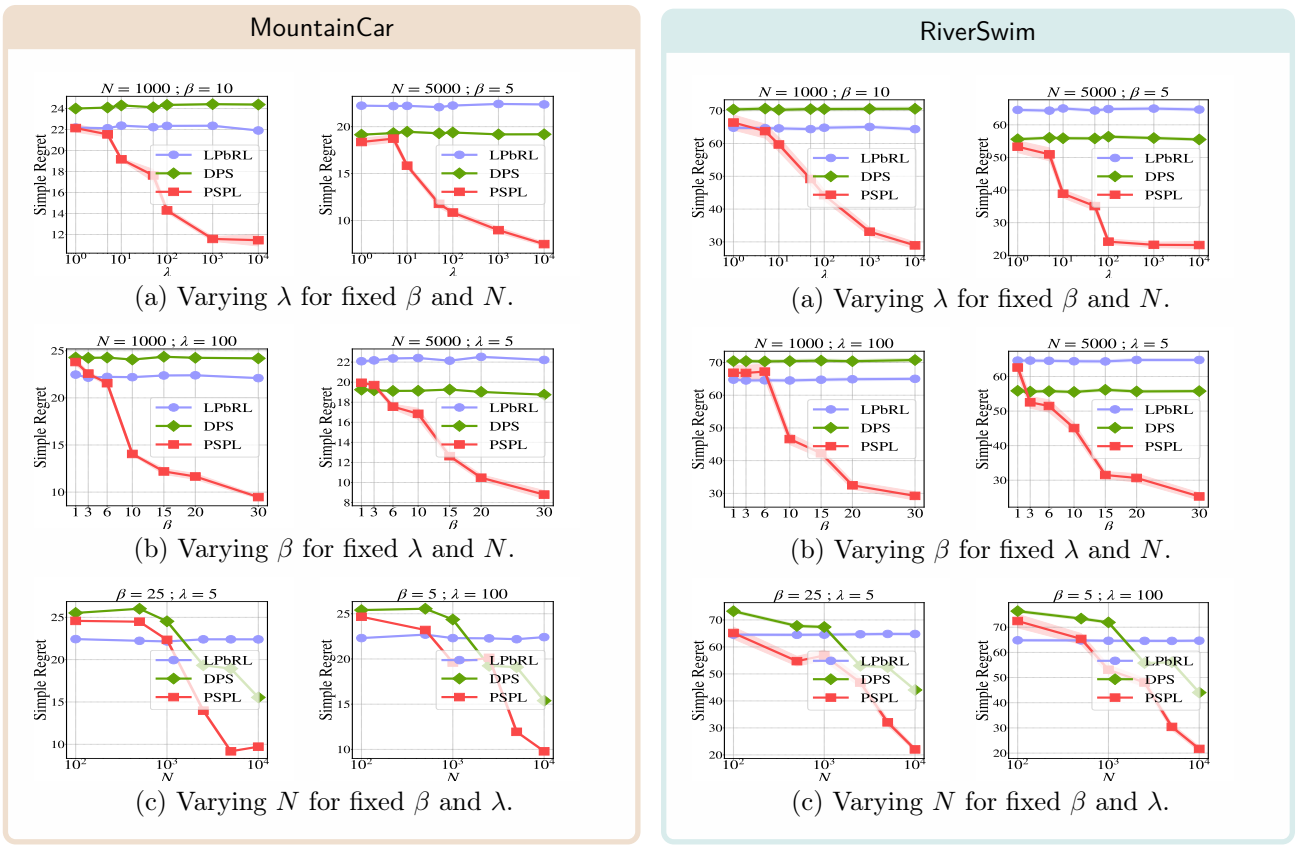
Sensitivity to specification errors. The Bootstrapped PSPL algorithm in Section 5 requires knowledge of rater's parameters λ, β. We study the sensitivity of PSPL's performance to mis-specification of these parameters. See Appendix A.4.1 for procedure to estimate λ, β in practice, and Appendix A.4.2 for ablation studies on robustness of PSPL to mis-specified parameters and mis-specified preference generation expert policies.

Results on Image Generation Tasks. We instantiate our framework on the Pick-a-Pic dataset of human preferences for text–to–image generation (Kirstain et al., 2023). The dataset contains over 500,000 examples and 35,000 distinct prompts. Each example contains a prompt, sequence generations of two images, and a corresponding preference label. We let each generation be a trajectory, so the dataset con-

tains trajectory preferences $\mathcal{D}_0 = (\tau_i^+, \tau_i^-, y_i)_{i=1}^N$ with $y_i = 1$ iff $\tau_i^+ \succ \tau_i^-$. Each trajectory $\tau = (p, z_{0:T})$ is the entire latent denoising chain $z_{0:T}$ of length T for prompt p sampled from some prompt distribution. See Appendix A.4.3 for the full experimental setup. As LPbRL involves optimization over confidence sets that grow exponentially in state-action space, we only present results of DPS and PSPL. See Figure 4 for example generations given some prompts, where we observe higher final reward (hence, lower simple regret) of PSPL as compared to DPS.

7 CONCLUSION

We have proposed the PSPL algorithm that incorporates offline preferences to learn an MDP with unknown rewards and transitions. We provide the first analysis to bridge the gap between offline preferences and online learning. We also introduced Bootstrapped PSPL, a computationally tractable extension, and our empirical results demonstrate superior performance of PSPL in challenging benchmarks as compared to baselines. Overall, we provide a promising foundation for further development in PbRL and RLHF space.



(a) Varying λ for fixed β and N.

(b) Varying β for fixed λ and N.

(c) Varying N for fixed β and λ.

Figure 2: PSPL with varying N, β, and λ in benchmark environments. Shaded region around mean line represents 1 standard deviation over 5 independent runs.

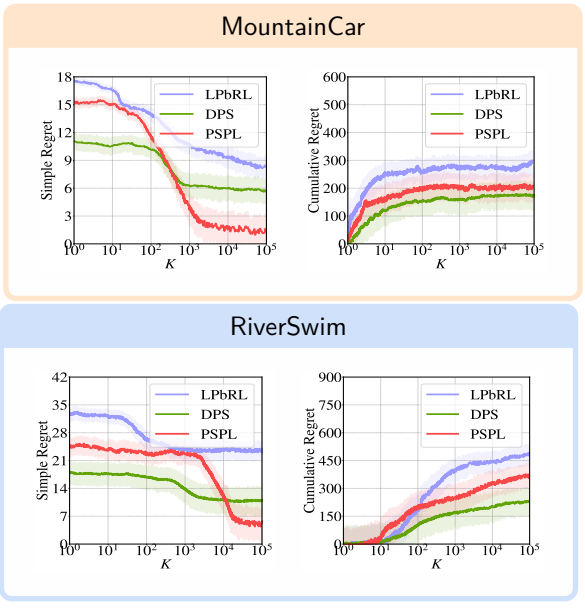


Figure 3: Simple and Cumulative Regret ($\div 10^3$) vs K plots. PSPL is run with $\lambda = 50, \beta = 10, N = 10^3$.

Figure 4: Sample image generations along with final image reward $\hat{r}_\theta(\cdot)$ over 5 independent runs.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Akhil Agnihotri, Rahul Jain, Deepak Ramachandran, and Zheng Wen. Online bandit learning with offline preference data. *arXiv preprint arXiv:2406.09574*, 2024.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

Mark Beliaev and Ramtin Pedarsani. Inverse reinforcement learning by estimating expertise of demonstrators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 15532–15540, 2025.

Mark Beliaev, Andy Shih, Stefano Ermon, Dorsa Sadigh, and Ramtin Pedarsani. Imitation learning by estimating expertise of demonstrators. In *International Conference on Machine Learning*, pages 1732–1748. PMLR, 2022.

Viktor Bengs, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 22(7):1–108, 2021.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444. URL `http://www.jstor.org/stable/2334029`.

Róbert Busa-Fekete and Eyke Hüllermeier. A survey of preference-based online learning with bandit algorithms. In *Algorithmic Learning Theory: 25th International Conference, ALT 2014, Bled, Slovenia, October 8-10, 2014. Proceedings 25*, pages 18–39. Springer, 2014.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Zhirui Chen and Vincent YF Tan. Order-optimal instance-dependent bounds for offline reinforcement learning with preference feedback. *arXiv preprint arXiv:2406.12205*, 2024.

Wang Chi Cheung and Lixing Lyu. Leveraging (biased) information: Multi-armed bandits with offline data. In *International Conference on Machine Learning*, pages 8286–8309. PMLR, 2024.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5713–5723, 2017.

Vikranth Dwaracherla, Zheng Wen, Ian Osband, Xiuyuan Lu, Seyed Mohammad Asghari, and Benjamin Van Roy. Ensembles for uncertainty estimation: Benefits of prior functions and bootstrapping. *Transactions on Machine Learning Research*, 2022.

LAION eV. laion2b-en-aesthetic (revision c2f3a74), 2025. URL `https://huggingface.co/datasets/laion/laion2B-en-aesthetic`.

Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, Aviv Tamar, et al. Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 8(5-6):359–483, 2015.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4447–4455. PMLR, 02–04 May 2024. URL `https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html`.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL `https://proceedings.mlr.press/v9/glorot10a.html`.

Javier González, Zhenwen Dai, Andreas Damianou, and Neil D Lawrence. Preferential bayesian optimization. In *International Conference on Machine Learning*, pages 1282–1291. PMLR, 2017.

Botao Hao, Rahul Jain, Dengwang Tang, and Zheng Wen. Bridging imitation and online reinforcement learning: An optimistic tale. *arXiv preprint arXiv:2303.11369*, 2023.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426, 1994.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, pages 4860–4869. PMLR, 2020.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 2023.

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

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Zihao Li, Xiang Ji, Minshuo Chen, and Mengdi Wang. Policy evaluation for reinforcement learning from human feedback: A sample complexity analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 2737–2745. PMLR, 2024.

Yi Liu, Gaurav Datta, Ellen Novoseller, and Daniel S Brown. Efficient preference-based reinforcement learning using learned dynamics models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2921–2928. IEEE, 2023.

Xiuyuan Lu and Benjamin Van Roy. Ensemble sampling. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf`.

Xiaofeng Mao, Yufeng Chen, Xiaojun Jia, Rong Zhang, Hui Xue, and Zhao Li. Context-aware robust fine-tuning. *International Journal of Computer Vision*, 132(5):1685–1700, 2024.

Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.

Katherine Metcalf, Miguel Sarabia, Natalie Mackraz, and Barry-John Theobald. Sample-efficient preference-based reinforcement learning with dynamics aware rewards. *arXiv preprint arXiv:2402.17975*, 2024.

Yifei Ming and Yixuan Li. How does fine-tuning impact out-of-distribution detection for vision-language models? *International Journal of Computer Vision*, 132(2):596–609, 2024.

Andrew William Moore. Efficient memory-based learning for robot control. Technical report, University of Cambridge, Computer Laboratory, 1990.

Ellen Novoseller, Yibing Wei, Yanan Sui, Yisong Yue, and Joel Burdick. Dueling posterior sampling for preference-based reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, pages 1029–1038. PMLR, 2020.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

Ian Osband, Benjamin Van Roy, Daniel J Russo, Zheng Wen, et al. Deep exploration via randomized value functions. *J. Mach. Learn. Res.*, 20(124):1–62, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision

processes: a thompson sampling approach. In *Proceedings of the International Conference on Neural Information Processing Systems*, 2017.

Chao Qin, Zheng Wen, Xiuyuan Lu, and Benjamin Van Roy. An analysis of ensemble sampling. *Advances in Neural Information Processing Systems*, 35:21602–21614, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

Dorsa Sadigh, Anca Dragan, Shankar Sastry, and Sanjit Seshia. *Active preference-based learning of reward functions*. escholarship.org, 2017.

Aadirupa Saha. Optimal algorithms for stochastic contextual preference bandits. *Advances in Neural Information Processing Systems*, 34:30050–30062, 2021.

Aadirupa Saha and Pierre Gaillard. Versatile dueling bandits: Best-of-both world analyses for learning from relative preferences. In *International Conference on Machine Learning*, pages 19011–19026. PMLR, 2022.

Aadirupa Saha, Aldo Pacchiano, and Jonathan Lee. Dueling rl: Reinforcement learning with trajectory preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 6263–6289. PMLR, 2023.

Antoine Scheid, Etienne Boursier, Alain Durmus, Michael I Jordan, Pierre Ménard, Eric Moulines, and Michal Valko. Optimal design for reward modeling in rlhf. *arXiv preprint arXiv:2410.17055*, 2024.

Lior Shani, Aviv Rosenberg, Asaf Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szpektor, et al. Multi-turn reinforcement learning from preference human feedback. *arXiv preprint arXiv:2405.14655*, 2024.

Satinder Singh, Diane Litman, Michael Kearns, and Marilyn Walker. Optimizing dialogue management with reinforcement learning: Experiments with the njfun system. *Journal of Artificial Intelligence Research*, 16:105–133, 2002.

Yuda Song, Yifei Zhou, Ayush Sekhari, J Andrew Bagnell, Akshay Krishnamurthy, and Wen Sun. Hybrid rl: Using both offline and online data can make rl efficient. *arXiv preprint arXiv:2210.06718*, 2022.

Alexander L. Strehl and Michael L. Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008. ISSN 0022-0000. doi: https://doi.org/10.1016/j.jcss.2007.08.009. URL `https://www.sciencedirect.com/science/article/pii/S0022000008000767`. Learning Theory 2005.

Gokul Swamy, David Wu, Sanjiban Choudhury, Drew Bagnell, and Steven Wu. Inverse reinforcement learning without reinforcement learning. In *International Conference on Machine Learning*, pages 33299–33318. PMLR, 2023.

Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *Algorithmic Learning Theory*, pages 770–805. PMLR, 2018.

Maegan Tucker, Ellen Novoseller, Claudia Kann, Yanan Sui, Yisong Yue, Joel W Burdick, and Aaron D Ames. Preference-based learning for exoskeleton gait optimization. In *2020 IEEE international conference on robotics and automation (ICRA)*, pages 2351–2357. IEEE, 2020.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Andrew Wagenmaker and Kevin Jamieson. Instance-dependent near-optimal policy identification in linear mdps via online experiment design. *arXiv preprint arXiv:2207.02575*, 2022.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, page 125, 2003.

Christian Wirth and Johannes Fürnkranz. Preference-based reinforcement learning: A preliminary survey. In *Proceedings of the ECML/PKDD-13 Workshop on Reinforcement Learning from Generalized Feedback: Beyond Numeric Rewards*, 2013.

Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46, 2017.

Wei Xiong, Hanze Dong, Chenlu Ye, Han Zhong, Nan Jiang, and Tong Zhang. Gibbs sampling from human feedback: A provable kl-constrained framework for rlhf. *arXiv preprint arXiv:2312.11456*, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Tengyu Xu, Yue Wang, Shaofeng Zou, and Yingbin Liang. Provably efficient offline reinforcement learning with trajectory-wise reward. *IEEE Transactions on Information Theory*, 2024.

Chenlu Ye, Wei Xiong, Yuheng Zhang, Nan Jiang, and Tong Zhang. A theoretical analysis of nash learning from human feedback under general kl-regularized preference. *arXiv preprint arXiv:2402.07314*, 2024.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312, 2019.

Yuexiang Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Shengbang Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, et al. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. *arXiv preprint arXiv:2405.10292*, 2024.

Ruiqi Zhang and Andrea Zanette. Policy finetuning in reinforcement learning via design of experiments using offline data. *Advances in Neural Information Processing Systems*, 36, 2024.

Yuheng Zhang, Dian Yu, Baolin Peng, Linfeng Song, Ye Tian, Mingyue Huo, Nan Jiang, Haitao Mi, and Dong Yu. Iterative nash policy optimization: Aligning llms with general preferences via no-regret learning. *arXiv preprint arXiv:2407.00617*, 2024a.

Zhilong Zhang, Yihao Sun, Junyin Ye, Tian-Shuo Liu, Jiaji Zhang, and Yang Yu. Flow to better: Offline preference-based reinforcement learning via preferred trajectory generation. In *The Twelfth International Conference on Learning Representations*, 2024b. URL `https://openreview.net/forum?id=EG68RSznLT`.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. Archer: Training language model agents via hierarchical multi-turn rl. *arXiv preprint arXiv:2402.19446*, 2024.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pages 43037–43067. PMLR, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes, please see Section 2.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes, please see Section 4.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. No, code will be released later.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes, please see Sections 3 and 4, and Appendix A.
   (b) Complete proofs of all theoretical results. Yes, please see Appendix A.
   (c) Clear explanations of any assumptions. Yes, please see Appendix A.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes, code will be released later. Please see Appendix A for instructions and experimental setup.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes, please see Appendix A.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes, please see Section 6 and Appendix A.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes, please see Appendix A.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes, please see Appendix A.
   (b) The license information of the assets, if applicable. Yes, please see Appendix A.
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable.

(d) Information about consent from data providers/curators. Yes, please see Appendix A.

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. Not Applicable.

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

Supplementary Material

A APPENDIX

Network Architecture for DPO Rafailov et al. (2024) and IPO Gheshlaghi Azar et al. (2024). We use a lightweight shared-trunk MLP with two hidden layers of 64 ReLU units each (Xavier initialization Glorot and Bengio (2010)), followed by a policy action head, outputting action logits for MountainCar (Moore, 1990) and RiverSwim (Strehl and Littman, 2008) environments. We train with Adam (learning rate 3×10^{-4} Kingma and Ba (2014)), and apply dropout (0.1) on the trunk to regularize. This architecture balances expressivity and sample efficiency on these control tasks.

DPO Rafailov et al. (2024) is an alternative approach to the RL paradigm, which avoids the training of a reward model altogether. The loss that DPO optimizes to obtain the optimal policy, given an empirical dataset $\mathcal{D} = \{y_w, y_l\}$ of the winning (preferred) y_w and losing (not preferred) y_l trajectories, as a function of the reference policy π_{ref} and regularization strength $\tau \in \mathbb{R}$, is given by:

$$\pi^{\star}_{\text{DPO}} = \operatorname*{argmin}_{\pi} \quad \mathbb{E}_{(y_w,y_l)\sim\mathcal{D}}\left[-\log\sigma\left(\tau\log\left(\frac{\pi(y_w)}{\pi(y_l)}\right) - \tau\log\left(\frac{\pi_{\text{ref}}(y_w)}{\pi_{\text{ref}}(y_l)}\right)\right)\right]$$

, where $\sigma(\cdot)$ denotes the sigmoid function.

IPO is an instance of the ΨPO algorithm Gheshlaghi Azar et al. (2024) . The loss function that IPO optimizes is given by,

$$\pi^{\star}_{\text{IPO}} = \operatorname*{argmin}_{\pi} \quad \mathbb{E}_{(y_w,y_l)\sim\mathcal{D}}\left[h_\pi(y_w,y_l) - \frac{1}{2\tau}\right]^2 \text{ where}, \; h_\pi(y,y') := \log\left(\frac{\pi(y)\pi_{\text{ref}}(y')}{\pi(y')\pi_{\text{ref}}(y)}\right).$$

A.1 Prior-dependent analysis

Lemma A.1 (Monotone Contraction)**.** PSPL *is monotone with respect to optimality of candidate policies* $\pi_k^{(0)}$ *and* $\pi_k^{(1)}$ *at any episode* $k \in [K]$ *i.e.* $\Pr(\pi_k^{(i)} \neq \pi^\star) \leqslant \Pr(\pi_1^{(i)} \neq \pi^\star)$ *for* $i \in \{0,1\} \quad \forall\, k \geqslant 1$.

Proof. Define $f(x) = x(1-x)$. f is a concave function. We have for any $i \in \{0,1\}$,

$$\begin{aligned}
\Pr(\pi_k^{(i)} \neq \pi^\star) &= \mathbb{E}\left[\Pr\left(\pi_k^{(i)} \neq \pi^\star \mid \mathcal{D}_k\right)\right] = \mathbb{E}\left[\sum_{\pi\in\Pi}\Pr(\pi_k^{(i)} = \pi, \pi^\star \neq \pi \mid \mathcal{D}_k)\right] \\
&= \mathbb{E}\left[\sum_{\pi\in\Pi} f\left(\Pr(\pi^\star = \pi \mid \mathcal{D}_k)\right)\right] \\
&= \mathbb{E}\left[\sum_{\pi\in\Pi} \mathop{\mathbb{E}}_{\mathcal{D}_1}\left[f\left(\Pr\left(\pi^\star = \pi \mid \mathcal{D}_k\right)\right)\right]\right] \leqslant \mathbb{E}\left[\sum_{\pi\in\Pi} f\left(\mathop{\mathbb{E}}_{\mathcal{D}_1}\left[\Pr\left(\pi^\star = \pi \mid \mathcal{D}_k\right)\right]\right)\right] \\
&= \mathbb{E}\left[\sum_{\pi\in\Pi} f\left(\Pr\left(\pi^\star = \pi \mid \mathcal{D}_1\right)\right)\right] = \mathbb{E}\left[\sum_{\pi\in\Pi}\Pr(\pi_1^{(i)} = \pi, \pi^\star \neq \pi \mid \mathcal{D}_1)\right] \\
&= \Pr(\pi_1^{(i)} \neq \pi^\star) \\
\Rightarrow \Pr(\pi_k^{(i)} \neq \pi^\star) &\leqslant \Pr(\pi_1^{(i)} \neq \pi^\star)
\end{aligned}$$

□

Lemma A.2. *For any policy π, let an event $E_1 := \left\{k, s, a : n_k(s,a) < \frac{1}{2}\sum_{j<k} w_j(s,a) - H\ln\left(\frac{SAH}{\delta'}\right)\right\}$ for some $\delta' \in (0,1)$, where $w_{h,j}(s,a)$ is the probability of visiting the (s,a) pair at time h of episode j under the chosen policy, and $w_j(s,a) = \sum_h w_{h,j}(s,a)$ is the sum of the probabilities of visiting the (s,a) pair in episode j. Then, $\Pr(E_1) \leqslant \delta' SAH = \delta$, where we set $\delta' = \frac{\delta}{SAH}$.*

Proof. Consider a fixed $s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$, and denote the state and action visited in the k^{th} episode at step h as $s_{k,h}$ and $a_{k,h}$ respectively. We define $\mathcal{F}_k$ to be the sigma-field induced by the first $k-1$ episodes. Let X_k be the indicator whether s, a was observed in episode k at time h. The probability $\Pr(s = s_{k,h}, a = a_{k,h} \mid s_{k,1}, \pi_k^{(i)})$ (for $i \in \{0,1\}$) of whether $X_k = 1$ is $\mathcal{F}_k$-measurable and hence we can apply Lemma F.4 from Dann et al. (2017) with $W = \ln\frac{SAH}{\delta'}$ and obtain that $\Pr(E_1) \leqslant SAH\delta'$ after summing over all statements for $h \in [H]$ and applying the union bound over s, a, h. □

Lemma A.3. *For all estimates $\widehat{\theta}_k^{(i)}, \widehat{\eta}_k^{(i)}$ $(i \in \{0,1\})$ of θ, η in any fixed episode $k \in [K]$, and for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, with probability at least $1 - 2\delta$ we have,*

$$\sum_{s' \in \mathcal{S}} \left| \mathbb{P}_{\theta,\eta}(s' \mid s,a) - \mathbb{P}_{\widehat{\theta}_k^{(i)}, \widehat{\eta}_k^{(i)}}(s' \mid s,a) \right| \leqslant \sqrt{\frac{4S\ln\left(\frac{2KSA}{\delta}\right)}{\sum_{j<k} w_j(s,a) - 2H\ln\left(\frac{SAH}{\delta}\right)}}$$

Proof. Weissman et al. (2003) gives the following high probability bound on the one norm of the maximum likelihood estimate using posterior sampling. In particular, with probability at least $1-\delta$, it holds that:

$$\sum_{s' \in \mathcal{S}} \left| \mathbb{P}_{\theta,\eta}(s' \mid s,a) - \mathbb{P}_{\widehat{\theta}_k^{(i)}, \widehat{\eta}_k^{(i)}}(s' \mid s,a) \right| \leqslant \sqrt{2S\ln\left(\frac{2KSA}{\delta}\right) / n_k(s,a)} \quad .$$

Then, using Lemma A.2, the proof is complete. □

Lemma A.4. *Under Algorithm 1 for large enough K, every reachable state-action pair is visited infinitely-often.*

Proof. The proof proceeds by assuming that there exists a state-action pair that is visited only finitely-many times. This assumption will lead to a contradiction [2]: once this state-action pair is no longer visited, the reward model posterior is no longer updated with respect to it. Then, Algorithm 1 is guaranteed to eventually sample a high enough reward for this state-action that the resultant policy will prioritize visiting it.

First we note that Algorithm 1 is guaranteed to reach at least one state-action pair infinitely often: given our problem's finite state and action spaces, at least one state-action pair must be visited infinitely-often during execution of Algorithm 1. If all state-actions are *not* visited infinitely-often, there must exist a state-action pair (s,a) such that s is visited infinitely-often, while (s,a) is not. Otherwise, if all actions are selected infinitely-often in all infinitely-visited states, the finitely-visited states are unreachable (in which case these states are irrelevant to the learning process and simple regret minimization, and can be ignored). Without loss of generality, we label this state-action pair (s,a) as $\widetilde{s}_1$. To reach a contradiction, it suffices to show that $\widetilde{s}_1$ is visited infinitely-often.

Let $\boldsymbol{r}_1$ be the reward vector with a reward of 1 in state-action pair $\widetilde{s}_1$ and rewards of zero elsewhere. Let $\pi_{pi}(\widetilde{\eta}, \boldsymbol{r}_1)$ be the policy that maximizes the expected number of visits to $\widetilde{s}_1$ under dynamics $\widetilde{\eta}$ and reward vector $\boldsymbol{r}_1$:

$$\pi_{pi}(\widetilde{\eta}, \boldsymbol{r}_1) = \text{argmax}_\pi V(\widetilde{\eta}, \boldsymbol{r}_1, \pi), \quad \text{where,} \quad V(\widetilde{\eta}, \boldsymbol{r}, \pi) = \mathop{\mathbb{E}}_{s_1 \sim \rho}\left[\sum_{t=1}^{H} \overline{r}(s_t, \pi(s_t,t)) \,\middle|\, s_{t+1} \sim \mathbb{P}_{\widetilde{\eta}}, \overline{\boldsymbol{r}} = \boldsymbol{r}\right].$$

where $V(\widetilde{\eta}, \boldsymbol{r}_1, \pi)$ is the expected total reward of a length-H trajectory under $\widetilde{\eta}, \boldsymbol{r}_1$, and π, or equivalently (by definition of $\boldsymbol{r}_1$), the expected number of visits to state-action $\widetilde{s}_1$.

We next show that there exists a $\kappa > 0$ such that $\Pr(\pi = \pi_{pi}(\widetilde{\eta}, \boldsymbol{r}_1)) > \kappa$ for all possible values of $\widetilde{\eta}$. That is, for any sampled parameters $\widetilde{\eta}$, the probability of selecting policy $\pi_{pi}(\widetilde{\eta}, \boldsymbol{r}_1)$ is uniformly lower-bounded, implying that Algorithm 1 must eventually select $\pi_{pi}(\widetilde{\eta}, \boldsymbol{r}_1)$.

[2] Note that in finite-horizon MDPs, the concept of visiting a state finitely-many times is not the same as that of a transient state in an infinite Markov chain, because: 1) due to a finite horizon, the state is resampled from the initial state distribution ρ every H timesteps, and 2) the policy—which determines which state-action pairs can be reached in an episode—is also resampled every H timesteps.

Let $\widetilde{r}_j$ be the sampled reward associated with state-action pair $\widetilde{s}_j$ in a particular Algorithm 1 episode, for each state-action $j \in \{1, \ldots, d\}$, with $d = SA$. We show that conditioned on $\widetilde{\eta}$, there exists $v > 0$ such that if $\widetilde{r}_1$ exceeds $\max\{v\widetilde{r}_2, v\widetilde{r}_3, \ldots, v\widetilde{r}_d\}$, then policy iteration returns the policy $\pi_{pi}(\widetilde{\eta}, \boldsymbol{r}_1)$, which is the policy maximizing the expected amount of time spent in $\widetilde{s}_1$. This can be seen by setting $v := \frac{H}{\kappa_1}$, where κ_1 is the expected number of visits to $\widetilde{s}_1$ under $\pi_{pi}(\widetilde{\eta}, \boldsymbol{r}_1)$. Under this definition of v, the event $\{\widetilde{r}_1 \geqslant \max\{v\widetilde{r}_2, v\widetilde{r}_3, \ldots, v\widetilde{r}_d\}\}$ is equivalent to $\{\widetilde{r}_1\kappa_1 \geqslant h \max\{\widetilde{r}_2, \widetilde{r}_3, \ldots, \widetilde{r}_d\}\}$; the latter inequality implies that given $\widetilde{\eta}$ and $\widetilde{\boldsymbol{r}}$, the expected reward accumulated solely in state-action $\widetilde{s}_1$ exceeds the reward gained by repeatedly (during all H time-steps) visiting the state-action pair in the set $\{\widetilde{s}_2, \ldots, \widetilde{s}_d\}$ having the highest sampled reward. Clearly, in this situation, policy iteration results in the policy $\pi_{pi}(\widetilde{\eta}, \boldsymbol{r}_1)$.

Next we show that $v = \frac{h}{\kappa_1}$ is continuous in the sampled dynamics $\widetilde{\eta}$ by showing that κ_1 is continuous in $\widetilde{\eta}$. Recall that κ_1 is defined as expected number of visits to $\widetilde{s}_1$ under $\pi_{pi}(\widetilde{\eta}, \boldsymbol{r}_1)$. This is equivalent to the expected reward for following $\pi_{pi}(\widetilde{\eta}, \boldsymbol{r}_1)$ under dynamics $\widetilde{\eta}$ and rewards $\boldsymbol{r}_1$:

$$\kappa_1 = V(\widetilde{\eta}, \boldsymbol{r}_1, \pi_{pi}(\widetilde{\eta}, \boldsymbol{r}_1)) = \max_{\pi} V(\widetilde{\eta}, \boldsymbol{r}_1, \pi). \tag{10}$$

The value of any policy π is continuous in the transition dynamics parameters, and so $V(\widetilde{\eta}, \boldsymbol{r}_1, \pi)$ is continuous in $\widetilde{\eta}$. The maximum in (10) is taken over the finite set of deterministic policies; because a maximum over a finite number of continuous functions is also continuous, κ_1 is continuous in $\widetilde{\eta}$.

Next, recall that a continuous function on a compact set achieves its maximum and minimum values on that set. The set of all possible dynamics parameters $\widetilde{\eta}$ is such that for each state-action pair j, $\sum_{k=1}^{S} p_{jk} = 1$ and $p_{jk} \geqslant 0 \, \forall \, k$; the set of all possible vectors $\widetilde{\eta}$ is clearly closed and bounded, and hence compact. Therefore, v achieves its maximum and minimum values on this set, and so for any $\widetilde{\eta}$, $v \in [v_{\min}, v_{\max}]$, where $v_{\min} > 0$ (v is nonnegative by definition, and $v = 0$ is impossible, as it would imply that $\widetilde{s}_1$ is unreachable).

Then, $\Pr(\pi = \pi_{pi}(\widetilde{\eta}, \boldsymbol{r}_1))$ can then be expressed in terms of v and the parameters of the reward posterior. Firstly,

$$\Pr(\pi = \pi_{pi}(\widetilde{\eta}, \boldsymbol{r}_1)) \geqslant \Pr(\widetilde{r}_1 > \max\{v\widetilde{r}_2, \ldots, v\widetilde{r}_d\}) \geqslant \prod_{j=2}^{d} \Pr(\widetilde{r}_1 > v\widetilde{r}_j) = \prod_{j=2}^{d} [1 - \Pr(\widetilde{r}_1 - v\widetilde{r}_j \leqslant 0)]$$

The posterior updates for the reward model are given by Equation (4), which is intractable to compute in closed form. Since we have $\theta_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$, we can use the result of Lemma 3 in Appendix of Novoseller et al. (2020). The remaining proof thereby follows, and as a result, there exists some $\kappa > 0$ such that $\Pr(\pi = \pi_{pi}(\widetilde{\eta}, \boldsymbol{r}_1)) \geqslant \kappa > 0$.

In consequence, Algorithm 1 is guaranteed to infinitely-often sample pairs $(\widetilde{\eta}, \pi)$ such that $\pi = \pi_{pi}(\widetilde{\eta}, \boldsymbol{r}_1)$. As a result, Algorithm 1 infinitely-often samples policies that prioritize reaching $\widetilde{s}_1$ as quickly as possible. Such a policy always takes action a in state s. Furthermore, because s is visited infinitely-often, either a) $p_0(s) > 0$ or b) the infinitely-visited state-action pairs include a path with a nonzero probability of reaching s. In case a), since the initial state distribution is fixed, the MDP will infinitely-often begin in state s under the policy $\pi = \pi_{pi}(\widetilde{\eta}, \boldsymbol{r}_1)$, and so $\widetilde{s}_1$ will be visited infinitely-often. In case b), due to Lemma 1 in Appendix of Novoseller et al. (2020), the transition dynamics parameters for state-actions along the path to s converge to their true values (intuitively, the algorithm knows how to reach s). In episodes with the policy $\pi = \pi_{pi}(\widetilde{\eta}, \boldsymbol{r}_1)$, Algorithm 1 is thus guaranteed to reach $\widetilde{s}_1$ infinitely-often. Since Algorithm 1 selects $\pi_{pi}(\widetilde{\eta}, \boldsymbol{r}_1)$ infinitely-often, it must reach $\widetilde{s}_1$ infinitely-often. This presents a contradiction, and so every state-action pair must be visited infinitely-often as the number of episodes tend to infinity. □

Lemma A.5. *For any confidence $\delta_1 \in (0, \frac{1}{3})$, let $\delta_2 \in (c, 1)$ with $c \in (0, 1)$, be the probability that any optimal policy estimate $\widehat{\pi}^\star$ constructed from the offline preference dataset $\mathcal{D}_0$ is ε-optimal with probability at least $(1 - \delta_2)$ i.e. $\Pr\left(\mathop{\mathbb{E}}_{s\sim\rho}\left[V^{\pi^\star}_{\theta,\eta,0}(s) - V^{\widehat{\pi}^\star}_{\theta,\eta,0}(s)\right] > \varepsilon\right) < \delta_2$. Then, the simple Bayesian regret of the learner Υ is upper bounded with probability of at least $1 - 3\delta_1$ by,*

$$\mathcal{SR}^{\Upsilon}_K(\pi^\star_{K+1}, \pi^\star) \leqslant \sqrt{\frac{10\delta_2 S^2 AH^3 \ln\left(\frac{2KSA}{\delta_1}\right) + 3\delta_2 SAH^2\varepsilon^2}{2K\left(1 + \ln\frac{SAH}{\delta_1}\right) - \ln\frac{SAH}{\delta_1}}} \tag{11}$$

Proof. Let $\Theta := (\theta, \eta)$ denote the unknown true parameters of the MDP $\mathcal{M}$, and let $\widehat{\Theta}_k^{(i)} := \left(\widehat{\theta}_k^{(i)}, \widehat{\eta}_k^{(i)}\right)$ denote the sampled reward and transition parameters at episode k, which are used to compute policy $\pi_k^{(i)}$ for $i \in \{0,1\}$. Let $J_\pi^{\widetilde{\Theta}} := \mathop{\mathbb{E}}_{\widetilde{\Theta}, \tau\sim\pi}[r(\tau)]$ denote the expected total reward for a trajectory sampled from policy π under some environment $\widetilde{\Theta} := \left(\widetilde{\theta}, \widetilde{\eta}\right)$. Then define for each $i \in \{0,1\}$,

$$Z_k(i) := J_{\pi^\star}^{\Theta} - J_{\pi_k^{(i)}}^{\Theta} - \varepsilon \quad ; \quad \widetilde{Z}_k(i) := J_{\pi_k^{(i)}}^{\widehat{\Theta}_k^{(i)}} - J_{\pi_k^{(i)}}^{\Theta} - \varepsilon \quad ; \quad I_k(i) := \mathbf{I}\left\{J_{\pi_k^{(i)}}^{\Theta} \neq J_{\pi^\star}^{\Theta} - \varepsilon\right\}$$

First, note that $Z_k(i) = Z_k(i)I_k(i)$ with probability 1. Then compute,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{D}_k}[Z_k(i) - \widetilde{Z}_k(i)I_k(i)] &= \mathop{\mathbb{E}}_{\mathcal{D}_k}\left[\left(Z_k(i) - \widetilde{Z}_k(i)\right)I_k(i)\right] = \mathop{\mathbb{E}}_{\mathcal{D}_k}\left[\left(J_{\pi^\star}^{\Theta} - J_{\pi_k^{(i)}}^{\widehat{\Theta}_k^{(i)}}\right)I_k(i)\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{D}_k}\left[J_{\pi^\star}^{\Theta}\mathbf{I}\left\{J_{\pi_k^{(i)}}^{\Theta} \neq J_{\pi^\star}^{\Theta} - \varepsilon\right\}\right] - \mathop{\mathbb{E}}_{\mathcal{D}_k}\left[J_{\pi_k^{(i)}}^{\widehat{\Theta}_k^{(i)}}\mathbf{I}\left\{J_{\pi_k^{(i)}}^{\Theta} \neq J_{\pi^\star}^{\Theta} - \varepsilon\right\}\right] = 0,
\end{aligned}$$

where the last equality is true since Θ and $\widehat{\Theta}_k^{(i)}$ are independently identically distributed given $\mathcal{D}_k$. Therefore, we can write the simple Bayesian regret upper bounded by $\mathbb{E}[\widetilde{Z}_{K+1}(0)I_{K+1}(0)]$. By Cauchy-Schwartz inequality, we have

$$\mathbb{E}[\widetilde{Z}_{K+1}(0)I_{K+1}(0)] \leqslant \sqrt{(\mathbb{E}[I_{K+1}(0)^2])\left(\mathbb{E}[\widetilde{Z}_{K+1}(0)^2]\right)}$$

Since $\pi^\star := \operatorname{argmax}_\pi J_\pi^{\Theta}$, using Lemma A.1 in conjunction with Appendix B of Hao et al. (2023), the first part can be bounded by

$$\mathbb{E}[I_{K+1}(0)^2] \leqslant \Pr\left(J_{\pi_{K+1}^{(0)}}^{\Theta} < J_{\pi^\star}^{\Theta} - \varepsilon\right) \leqslant \max_{i\in\{0,1\}} \Pr\left(J_{\pi_1^{(i)}}^{\Theta} < J_{\pi^\star}^{\Theta} - \varepsilon\right) \leqslant \delta_2.$$

Let $\mathcal{T}_{\pi_h}^{\Theta}$ be the Bellman operator at time h defined by $\mathcal{T}_{\pi_h}^{\Theta}V_{h+1}(s) := r_\theta(s_h, a_h) + \sum_{s'\in\mathcal{S}} V_{h+1}(s')\mathbb{P}_\eta(s'|s_h, \pi(s_h))$ and $\mathcal{T}_{\pi_H}^{\Theta}V_{H+1}(s) = 0$ for all $s \in \mathcal{S}$. Using Equation (6) of Osband et al. (2013) (also see Lemma A.14 of Zhang and Zanette (2024)), we have

$$\widetilde{Z}_{K+1}(0) = \mathop{\mathbb{E}}_{\Theta, \widehat{\Theta}_{K+1}^{(0)}}\left[\sum_{h=1}^{H}\left[\mathcal{T}_{\pi_h^{K+1}}^{\widehat{\Theta}_{K+1}^{(0)}}V_{h+1}^{\widehat{\Theta}_{K+1}^{(0)}}(s_{K+1,h}) - \mathcal{T}_{\pi_h^{K+1}}^{\Theta}V_{h+1}^{\widehat{\Theta}_{K+1}^{(0)}}(s_{K+1,h})\right]\right]$$

Recall that the instantaneous reward satisfies $r_\theta(\cdot,\cdot) \in [0,1]$. So we have with $a_{K+1,h} := \pi_{K+1}(s_{K+1,h})$,

$$\left|\mathcal{T}_{\pi_h^{K+1}}^{\widehat{\Theta}_{K+1}^{(0)}}V_{h+1}^{\widehat{\Theta}_{K+1}^{(0)}}(s_{K+1,h}) - \mathcal{T}_{\pi_h^{K+1}}^{\Theta}V_{h+1}^{\widehat{\Theta}_{K+1}^{(0)}}(s_{K+1,h})\right| \leqslant H\|\mathbb{P}_{\widehat{\eta}_{K+1}^{(0)}}(\cdot \mid s_{K+1,h}, a_{K+1,h}) - \mathbb{P}_\eta(\cdot \mid s_{K+1,h}, a_{K+1,h})\|_1^1$$

Therefore,

$$\begin{aligned}
\mathbb{E}[\widetilde{Z}_{K+1}(0)^2] &\leqslant H\mathbb{E}\left[\sum_{h=1}^{H}\left[\mathcal{T}_{\pi_h^{K+1}}^{\widehat{\Theta}_{K+1}^{(0)}}V_{h+1}^{\widehat{\Theta}_{K+1}^{(0)}}(s_{K+1,h}) - \mathcal{T}_{\pi_h^{K+1}}^{\Theta}V_{h+1}^{\widehat{\Theta}_{K+1}^{(0)}}(s_{K+1,h})\right]^2\right] && \text{(Cauchy-Schwartz)} \\
&\leqslant H^3\mathbb{E}\left[\sum_{h=1}^{H}\|\mathbb{P}_{\widehat{\eta}_{K+1}^{(0)}}(\cdot \mid s_{K+1,h}, a_{K+1,h}) - \mathbb{P}_\eta(\cdot \mid s_{K+1,h}, a_{K+1,h})\|_1^2\right] \\
&\leqslant 5H^3\mathbb{E}\left[\sum_{h=1}^{H}\frac{4S\ln\left(\frac{2KSA}{\delta_1}\right)}{\sum_{j<K+1} w_j(s_{K+1,h}, a_{K+1,h}) - 2H\ln\frac{SAH}{\delta_1}}\right]
\end{aligned}$$

, where the last line holds with probability $(1-2\delta_1)$ from Lemma A.3 and Section H.3 in Zanette and Brunskill (2019). Now it remains to bound $\sum_{j<K+1} w_j(s_{K+1,h}, a_{K+1,h})$, which is essentially the sum of probabilities of visiting the pair $(s_{K+1,h}, a_{K+1,h})$ before the $(K+1)^{\text{th}}$ episode. We know from Appendix G of Zanette and Brunskill (2019) (also see Appendix B of Jin et al. (2020)) that with probability at least $(1-\delta')$, for any state-action (s,a) in any episode e, we have $\frac{1}{4}w_e(s,a) \geqslant H \ln \frac{SAH}{\delta'} + H$. Using this, we have with probability at least $(1-3\delta_1)$,

$$\begin{aligned}
\mathbb{E}[\widetilde{Z}_{K+1}(0)^2] &\leqslant 5H^3\mathbb{E}\left[\frac{4SH\ln\left(\frac{2KSA}{\delta_1}\right)+\varepsilon^2}{4HK\left(1+\ln\frac{SAH}{\delta_1}\right)-2H\ln\frac{SAH}{\delta_1}}\right] \\
&\leqslant \frac{20S^2AH^4\ln\left(\frac{2KSA}{\delta_1}\right)+5SAH^3\varepsilon^2}{4HK\left(1+\ln\frac{SAH}{\delta_1}\right)-2H\ln\frac{SAH}{\delta_1}} \\
&\leqslant \frac{10S^2AH^3\ln\left(\frac{2KSA}{\delta_1}\right)+3SAH^2\varepsilon^2}{2K\left(1+\ln\frac{SAH}{\delta_1}\right)-\ln\frac{SAH}{\delta_1}} \,.
\end{aligned}$$

Putting it all together, we have with probability at least $(1-3\delta_1)$,

$$\mathfrak{SR}_K^{\Upsilon}(\pi^\star_{K+1}, \pi^\star) \leqslant \sqrt{\frac{10\delta_2 S^2AH^3\ln\left(\frac{2KSA}{\delta_1}\right)+3\delta_2 SAH^2\varepsilon^2}{2K\left(1+\ln\frac{SAH}{\delta_1}\right)-\ln\frac{SAH}{\delta}}}$$

□

Theorem A.6. *For any confidence $\delta_1 \in (0, \frac{1}{3})$ and offline preference dataset size $N > 2$, the simple Bayesian regret of the learner Υ is upper bounded with probability of at least $1-3\delta_1$ by,*

$$\begin{aligned}
\mathfrak{SR}_K^{\Upsilon}(\pi^\star_{K+1}, \pi^\star) &\leqslant \sqrt{\frac{20\delta_2 S^2AH^3\ln\left(\frac{2KSA}{\delta_1}\right)}{2K\left(1+\ln\frac{SAH}{\delta_1}\right)-\ln\frac{SAH}{\delta_1}}} \quad , \textit{where} \\
\delta_2 &= 2\exp\left(-N\left(1+\gamma_{\beta,\lambda,N}\right)^2\right)+\exp\left(-\frac{N}{4}(1-\gamma_{\beta,\lambda,N})^3\right)
\end{aligned} \tag{12}$$

Proof. Define an event $\mathcal{E}_n = \left\{\overline{Y}_n \neq \text{argmax}_{i\in\{0,1\}}\, g_{\beta,\vartheta}\left(\overline{\tau}_n^{(i)}\right)\right\}$, i.e. at the n-th index of the offline preference dataset, the rater preferred the suboptimal trajectory (wrt trajectory score $g(\cdot)$). Given the optimal trajectory parity at index n as $\overline{Y}^\star_n = \text{argmax}_{i\in\{0,1\}}\, g_{\beta,\vartheta}(\overline{\tau}_n^{(i)})$, we have,

$$
\begin{aligned}
\Pr(\mathcal{E}_n \,|\, \beta, \vartheta) &= 1 - \Pr(\mathcal{E}_n^c \,|\, \beta, \vartheta) \\
&= 1 - \frac{g_{\beta,\vartheta}\left(\overline{\tau}_n^{\overline{Y}^\star_n}\right)}{g_{\beta,\vartheta}\left(\overline{\tau}_n^{\overline{Y}^\star_n}\right) + g_{\beta,\vartheta}\left(\overline{\tau}_n^{(1-\overline{Y}^\star_n)}\right)} \\
&= 1 - \frac{1}{1 + \exp\left(\beta\left\langle \phi\left(\overline{\tau}_n^{(\overline{Y}^\star_n)}\right) - \phi\left(\overline{\tau}_n^{(1-\overline{Y}^\star_n)}\right), -\vartheta\right\rangle\right)} \\
&= 1 - \frac{1}{1 + \exp\left(\beta\left\langle \phi\left(\overline{\tau}_n^{(\overline{Y}^\star_n)}\right) - \phi\left(\overline{\tau}_n^{(1-\overline{Y}^\star_n)}\right), \theta - \vartheta\right\rangle - \beta\left\langle \phi\left(\overline{\tau}_n^{(\overline{Y}^\star_n)}\right) - \phi\left(\overline{\tau}_n^{(1-\overline{Y}^\star_n)}\right), \theta\right\rangle\right)} \\
&\leqslant 1 - \frac{1}{1 + \exp\left(\beta\left\|\phi\left(\overline{\tau}_n^{(\overline{Y}^\star_n)}\right) - \phi\left(\overline{\tau}_n^{(1-\overline{Y}^\star_n)}\right)\right\|_1 \|\theta - \vartheta\|_\infty - \beta\left\langle \phi\left(\overline{\tau}_n^{(\overline{Y}^\star_n)}\right) - \phi\left(\overline{\tau}_n^{(1-\overline{Y}^\star_n)}\right), \theta\right\rangle\right)} \\
&\leqslant 1 - \frac{1}{1 + \underbrace{\exp\left(\beta B\|\vartheta - \theta\|_\infty - \beta\left\langle \phi\left(\overline{\tau}_n^{(\overline{Y}^\star_n)}\right) - \phi\left(\overline{\tau}_n^{(1-\overline{Y}^\star_n)}\right), \theta\right\rangle\right)}_{\clubsuit}}
\end{aligned}
$$

, where the last two lines use Hölder's inequality, and bounded trajectory map assumption respectively. Since $\vartheta - \theta \sim \mathcal{N}(0, \mathbf{I}_d/\lambda^2)$, using the Dvoretzky–Kiefer–Wolfowitz inequality bound (Massart, 1990; Vershynin, 2010) implies

$$
\Pr\left(\|\vartheta - \theta\|_\infty \geqslant t\right) \leqslant 2d^{1/2} \exp\left(-\frac{t^2\lambda^2}{2}\right).
$$

Set $t = \sqrt{2\ln(2d^{1/2}N)}/\lambda$ and define an event $\mathcal{E}_{\theta,\vartheta} := \{\|\vartheta - \theta\|_\infty \leqslant \sqrt{2\ln(2d^{1/2}N)}/\lambda\}$ such that $P(\mathcal{E}^c_{\theta,\vartheta}) \leqslant 1/N$. We apply Union Bound on the $\clubsuit$ term, and decompose the entire right hand side term as:

$$
\begin{aligned}
\Pr(\mathcal{E}_n \,|\, \beta, \vartheta) &\leqslant \frac{1}{1 + \exp\left(\beta B\sqrt{2\ln(2d^{1/2}N)}/\lambda + \beta\Delta_{\min}\right)} + \frac{1}{N} \\
&\leqslant \exp\left(-\beta B\sqrt{2\ln(2d^{1/2}N)}/\lambda - \beta\Delta_{\min}\right) + \frac{1}{N} := \gamma_{\beta,\lambda,N}\,.
\end{aligned} \tag{13}
$$

Now we need to provide conditions on the rater's competence, in terms of (λ, β) to be a valid expert i.e. for $\gamma_{\beta,\lambda,N} \in (0,1)$. Let $k_1 = \beta B, k_2 = \frac{2\ln\left(2d^{1/2}\right)}{\lambda^2}, k_3 = 2/\lambda^2$, and $k_4 = \beta\Delta_{\min}$. We then have,

$$
\begin{aligned}
\exp\left(-k_1\sqrt{k_2 + k_3\ln N} - k_4\right) + \frac{1}{N} &< 1 \\
-k_1\sqrt{k_2 + k_3\ln N} - k_4 &< \ln N \\
(\ln N)^2 + (2k_4 - k_1^2k_3)\ln N + (k_4^2 - k_1^2k_2) &> 0
\end{aligned}
$$

The above is a quadratic inequality that holds for all $N > 1$ if β is large enough i.e.

$$
\text{If } \beta > \frac{2\ln\left(2d^{1/2}\right)}{|B\lambda^2 - 2\Delta_{\min}|}, \text{then, } \gamma_{\beta,\lambda,N} \in (0,1)\,.
$$

Now, since we have Lemma A.4, the remaining argument shows to find a separation of probability between two types of states and time index pairs under the rater's preference, parameterized by β and λ, and the offline dataset, characterized by its size N: the ones that are probable under the optimal policy $\pi^\star$, and the ones that are not. We have two cases:

Case I. $p_h^{\pi^\star}(s) > 0$

In this case, we want the rater to prefer trajectories that are most likely to occur under the optimal policy $\pi^\star$. Given $p_h^{\pi^\star}(s) > 0$, we now lower bound the probability of the state s having $c_h(s) \equiv \sum_{a\in\mathcal{A}} c_h(s,a) > 0$ i.e.

$$
\begin{aligned}
\Pr(c_h(s) > 0) &= \Pr\left(\sum_a w_h(s,a) > \sum_a l_h(s,a)\right) = 1 - \Pr\left(\sum_a w_h(s,a) \leqslant \sum_a l_h(s,a)\right) \\
&= 1 - \sum_{t=0}^{H} \Pr\left(\sum_a w_h(s,a) \leqslant t\right) \cdot \Pr\left(\sum_a l_h(s,a) = t\right) \\
&\geqslant 1 - \sum_{t=1}^{H+1} \Pr\left(\sum_a w_h(s,a) < t\right) \\
&\geqslant 1 - \gamma_{\beta,\lambda,N}/(1-\gamma_{\beta,\lambda,N})
\end{aligned} \tag{14}
$$

, where the last step uses Equation (13).

Case II. $p_h^{\pi^\star}(s) = 0$

In this case, we wish to upper bound the probability of the rater preferring trajectories (and hence, states) which are unlikely to occur under the optimal policy $\pi^\star$. If for some state s we have $p_h^{\pi^\star}(s) = 0$ but also this state $s \in \overline{\tau}_n^{(1-\overline{Y}^\star_n)}$, this means the rater preferred the suboptimal trajectory. Similar to the proof above we conclude that,

$$
\Pr(c_h(s) > 0) = \Pr\left(\bigcup_{n=1}^{N} \{\mathcal{E}_n\} \,|\, \beta, \vartheta\right) \leqslant \sum_{n=1}^{N} \Pr_\theta(\mathcal{E}_n \,|\, \beta, \vartheta) \leqslant \gamma_{\beta,\lambda,N}/(1-\gamma_{\beta,\lambda,N}) \tag{15}
$$

The above argument shows that there's a probability gap between parity of states and time index pairs under the rater's preference model: the ones that are probable under the optimal policy $\pi^\star$ and the ones that are not i.e. we have shown that the states which are *more* likely to be visited under $\pi^\star$ have a lower bound on the probability of being part of preferred trajectories, and also the states which are *less* likely to be visited by $\pi^\star$ have an upper bound on the probability of being part of preferred trajectories by the rater. Using this decomposition, we will show that when the rater tends to an expert ($\beta \to \infty, \lambda \to \infty$) and size N of the offline preference dataset $\mathcal{D}_0$ is large, we can distinguish the two types of state and time index pairs through their net counts in $\mathcal{D}_0$. This will allow us to construct an ε-optimal estimate of $\pi^\star$ with probability at least $(1-\delta_2)$. Noticing the structure of Equation (6), we see that the upper bound is minimized when $\varepsilon \to 0$, i.e. we construct an *optimal* estimate of the optimal policy from the offline preference dataset. We now upper bound the probability that the estimated optimal policy $\widehat{\pi}^\star$ is *not* the optimal policy.

Let $\widehat{\pi}^\star$ be the optimal estimator of $\pi^\star$ constructed with probability at least $(1-\delta_2)$. Based on the separability of states and time index pairs, we have four possible events for each $(s,h) \in \mathcal{S} \times [H]$ pair. For $\delta = (1-\gamma_{\beta,\lambda,N})/2$, we have,

1. $E_1 := \{p_h^{\pi^\star}(s) > 0 \text{ and } c_h(s) < \delta N\}$;
2. $E_2 := \{p_h^{\pi^\star}(s) > 0 \text{ and } c_h(s) \geqslant \delta N$, but $\pi_h^\star(s) = a_h^\star(s) \neq \operatorname{argmax}_a c_h(s,a) = \widehat{\pi}_h^\star(s)\}$.
3. $E_3 := \{p_h^{\pi^\star}(s) = 0 \text{ and } c_h(s) \geqslant \delta N\}$;
4. $E_4 := \{p_h^{\pi^\star}(s) = 0 \text{ and } c_h(s) < \delta N\}$;

Denoting the the event to occur with high probability as $\mathcal{T}$ i.e. $\mathcal{T} := \{r_\theta(s_h, \pi_h^\star(s_h)) - r_\theta(s_h, \widehat{\pi}_h^\star(s_h)) \leqslant 0\}$. If we can show that $\Pr(\mathcal{T} \,|\, E_i) > 1-\delta_2$ for $i \in \{1,2,3,4\}$, then union bound implies that $\widehat{\pi}^\star$ is not optimal with probability at most δ_2.

1. **Under the event E_1.** Let $b \sim \text{Bin}(T,q)$ denote a binomial random variable with parameters $T \in \mathbb{N}$ and $q \in [0,1]$. Notice that the each $c_h(s)$ is the difference of two binomial random variables $b_1 \sim \text{Bin}(N, 1-\gamma_{\beta,\lambda,N})$ and $b_2 \sim \text{Bin}(N, \gamma_{\beta,\lambda,N})$. This implies that $c_h(s) + N \sim \text{Bin}(2N, 1-\gamma_{\beta,\lambda,N})$. We then have,

$$
\begin{aligned}
\Pr(\mathcal{T}^c \,|\, E_1) &\leqslant \Pr(c_h(s) < \delta N) \leqslant \Pr\left(\text{Bin}(2N, 1-\gamma_{\beta,\lambda,N}) < (1+\delta)N\right) \\
&\leqslant \exp\left(-4N\left(\delta + \gamma_{\beta,\lambda,N}\right)^2\right) \leqslant \exp\left(-N\left(1+\gamma_{\beta,\lambda,N}\right)^2\right)
\end{aligned}
$$

2. **Under the event E_2.** Given Equation (14), we have that,

$$\begin{aligned}
\Pr(\mathcal{T}^c \mid E_2) &\leqslant \Pr\left(\operatorname*{argmax}_{a\in U_h^W(s)} Q^{\pi^\star}_{\theta,\eta,h}(s,a) \neq \operatorname*{argmax}_{a\in U_h^W(s)} Q^{\pi^\star}_{\theta,\eta,h}(s,a) \mid E_2\right) \\
&\leqslant \Pr\left(c_h\left(s, \operatorname*{argmax}_{a\in U_h^W(s)} Q^{\pi^\star}_{\theta,\eta,h}(s,a)\right) \leqslant c_h(s)/2 \mid E_2\right) \\
&\leqslant \Pr\left(\mathrm{Bin}(c_h(s), 1-\gamma_{\beta,\lambda,N}) \geqslant c_h(s)/2 \mid E_2\right) \\
&\leqslant \left[\exp\left(-2c_h(s)(1-\gamma_{\beta,\lambda,N}-c_h(s)/2)^2\right)\right]|_{c_h(s)=\delta N} \\
&\leqslant \exp\left(-\frac{N}{4}(1-\gamma_{\beta,\lambda,N})^3\right)
\end{aligned}$$

3. **Under the event E_3.** Similar to event E_1, we have,

$$\begin{aligned}
\Pr(\mathcal{T}^c \mid E_3) &\leqslant \Pr\left(\mathrm{Bin}(2N, 1-\gamma_{\beta,\lambda,N}) > (1+\delta)N\right) \\
&\leqslant \exp\left(-4N\left(\delta+\gamma_{\beta,\lambda,N}\right)^2\right) \leqslant \exp\left(-N\left(1+\gamma_{\beta,\lambda,N}\right)^2\right)
\end{aligned}$$

4. **Under the event E_4.** Under this event, notice that conditioned on θ, we have

$$\begin{aligned}
r_\theta(s_h, \pi^\star_h(s_h)) - r_\theta(s_h, \widehat{\pi}^\star_h(s_h)) &= \mathbb{E}\left[\mathop{\mathbb{E}}_{\theta}\left[r_\theta(s_h, \pi^\star_h(s_h)) - r_\theta(s_h, \widehat{\pi}^\star_h(s))\right]\right] \\
&= \mathop{\mathbb{E}}_{\mathring{a}\sim\mathcal{A}}\left[\mathop{\mathbb{E}}_{\theta}\left[r_\theta(s_h, \mathring{a}) - r_\theta(s_h, \mathring{a})\right]\right] = 0
\end{aligned}$$

This means that $\Pr(\mathcal{T} \mid E_4)$ is a non-failure event that occurs with probability 1 i.e. $\Pr(\mathcal{T}^c \mid E_4) = 0$.

Combining all of the above, we have for $N > 2$,

$$\begin{aligned}
\Pr(\mathcal{T}) &\geqslant 1 - \left(2\exp\left(-N\left(1+\gamma_{\beta,\lambda,N}\right)^2\right) + \exp\left(-\frac{N}{4}(1-\gamma_{\beta,\lambda,N})^3\right)\right) \\
&\geqslant 1-\delta_2
\end{aligned}$$

Using Lemma 4.1, the proof is complete.

□

A.2 Results for Bandits

For the bandit setting we let the action set be $\mathcal{A} \subseteq \mathbb{R}^d$ with number of arms be $A = |\mathcal{A}|$, online episodes (rounds) be K, and offline dataset $\mathcal{D}_0 = \left\{\left(\overline{a}_n^{(0)}, \overline{a}_n^{(1)}, Y_n\right)\right\}_{n=1}^N$ of size N, where $\overline{a}_n^{(0)}, \overline{a}_n^{(1)} \in \mathcal{A}$. Also let $\mu_{\min}(\cdot) \in (0,1)$ be the minimum action sampling distribution during construction of $\mathcal{D}_0$.

Letting $\Delta_{\min} = \min_{n\in[N]}\left|r_\theta(\overline{a}_n^{(0)}) - r_\theta(\overline{a}_n^{(1)})\right|$, where θ is the underlying reward model of the environment with $r_\theta(\cdot) : \mathbb{R}^d \to \mathbb{R}$, and $\gamma_{\beta,\lambda,N}$ to be error upper bound of the rater's preference (similar to Lemma 4.3), we have the following result for simple regret of the learner Υ, where simple regret is defined as $\mathcal{SR}_K^\Upsilon(\pi^\star_{K+1}, \pi^\star) = r_\theta(a^\star) - r_\theta(a^\star_{K+1})$, where $\pi^\star$ is the optimal policy that picks the optimal action $a^\star = \operatorname{argmax}_{a\in\mathcal{A}} r_\theta(a)$, and $\pi^\star_{K+1}$ is the policy of the learner after K online rounds. We shall use $A^\star$ and $a^\star$ interchangebly to refer to the optimal action.

Analogous to the winning and undecided subsets of Section 4, we construct an *information* subset of $\mathcal{A}$, call it $\mathcal{U}_{\mathcal{D}_0}$ such that $\Pr(a^\star \in \mathcal{U}_{\mathcal{D}_0}) \geqslant 1-\epsilon$, where $\epsilon \in (0,1)$ is the *error* probability. As an algorithmic choice, we let $\mathcal{U}_{\mathcal{D}_0}$ consist of actions that have been preferred to *at least* once in the offline preference dataset and of actions that do not appear in $\mathcal{D}_0$. Given this construction, we have the following lemma, which has similarities to Appendix A of Agnihotri et al. (2024).

Theorem A.7. *Given the construction of $\mathcal{U}_{\mathcal{D}_0}$ above, we have* $\Pr(a^\star \in \mathcal{U}_{\mathcal{D}_0}) \geqslant 1 - f_1$, *where*

$$f_1 := \left(1 - \frac{1}{1 + \exp\left(\beta\big(\min(1,\Delta) + \alpha_2 - \alpha_1^{\Delta}\big)\right)}\right)^N + (1-\mu_{min})^{2N} + \frac{1}{K} \tag{16}$$
$$\text{and,} \quad \Delta := \ln(K\beta)/\beta\,,\ \alpha_1^{\Delta} := A\min(1,\Delta),\ \text{and}\ \alpha_2 := \lambda^{-1}\sqrt{2\ln(2d^{1/2}K)}$$

Proof. We construct $\mathcal{U}_{\mathcal{D}_0}$ as a *set* of actions that have been preferred to *at least* once in the offline dataset $\mathcal{D}_0$ and of actions that do not appear in the $\mathcal{D}_0$. Thus, $\mathcal{U}_{\mathcal{D}_0}$ contains at most A actions.

Now, we consider the formulation below. Recall that $\overline{A}_n^{(0)}$ and $\overline{A}_n^{(1)}$ are i.i.d. sampled from the action set and each datapoint in the dataset $\mathcal{D}_0^i$, conditioned on ϑ, β, is independent of $\mathcal{D}_0^j$ for $i \neq j$. Now,

$$\begin{aligned}
P\left(A^\star \notin \mathcal{U}_{\mathcal{D}_0}\right) &\leqslant P(A^\star \text{ has lost all comparisons in } \mathcal{D}_0) + P(A^\star \text{ is not present in } \mathcal{D}_0) \\
&\leqslant \mathbb{E}\Big[\prod_{n=1}^{N} \frac{\exp\left(\beta\langle a_n, \vartheta\rangle\right)}{\exp\left(\beta\langle a_n, \vartheta\rangle\right) + \exp\left(\beta\langle A^\star, \vartheta\rangle\right)} + (1-\mu_{\min})^{2N}\Big] \\
&\leqslant \mathbb{E}\Big[\prod_{n=1}^{N}\left(1 - \frac{\exp\left(\beta\langle A^\star, \vartheta\rangle\right)}{\exp\left(\beta\langle a_n, \vartheta\rangle\right) + \exp\left(\beta\langle A^\star, \vartheta\rangle\right)}\right)\Big] + (1-\mu_{\min})^{2N} \\
&\leqslant \mathbb{E}\Big[\prod_{n=1}^{N}\Big(1 - \underbrace{\frac{1}{1 + \exp\left(-\beta\langle A^\star - a_n, \vartheta\rangle\right)}}_{\clubsuit}\Big)\Big] + (1-\mu_{\min})^{2N}
\end{aligned} \tag{17}$$

, where $A^\star$ is a function of θ and thus a random variable as well. Looking closely at the term $\clubsuit$ above, it can be written as $P(Y_n = A^\star \mid \vartheta)$. We now analyze this term.

$$\begin{aligned}
P(Y_n = A^\star \mid \vartheta) &= \frac{1}{1 + \exp\left(-\beta\langle A^\star - a_n, \vartheta\rangle\right)} \\
&= \left(1 + \exp\left(\beta\langle A^\star - a_n, \theta - \vartheta\rangle - \beta\langle A^\star - a_n, \theta\rangle\right)\right)^{-1} \\
&\geqslant \left(1 + \exp\left(\beta\|A^\star - a_n\|_1\|\theta - \vartheta\|_\infty - \beta\langle A^\star - a_n, \theta\rangle\right)\right)^{-1} && \text{(Hölder's inequality)} \\
&\geqslant \left(1 + \exp\left(\beta\|\vartheta - \theta\|_\infty - \beta\langle A^\star - a_n, \theta\rangle\right)\right)^{-1} && (\|A^\star - a_n\|_1 \leqslant 1 \ \ \forall\ a_n \in \mathcal{A})
\end{aligned}$$

Since $\vartheta - \theta \sim N(0, \mathbf{I}_d/\lambda^2)$, using the Dvoretzky–Kiefer–Wolfowitz inequality bound (Massart, 1990; Vershynin, 2010) implies

$$P\left(\|\vartheta - \theta\|_\infty \geqslant t\right) \leqslant 2d^{1/2}\exp\left(-\frac{t^2\lambda^2}{2}\right).$$

Set $t = \sqrt{2\ln(2d^{1/2}K)}/\lambda$ and define an event $\mathcal{E}_1 := \{\|\vartheta - \theta\|_\infty \leqslant \sqrt{2\ln(2d^{1/2}K)}/\lambda\}$ such that $P(\mathcal{E}_1^c) \leqslant 1/K$. We decompose Equation (17) using Union Bound as:

$$\begin{aligned}
P\left(A^\star \notin \mathcal{U}_{\mathcal{D}_0}\right) &\leqslant \mathbb{E}\left[\prod_{n=1}^{N}\left(1 - P\left(Y_n = A^\star \mid \theta, \vartheta\right)\right)\mathbb{I}_{\mathcal{E}_1}\right] + P(\mathcal{E}_1^c) + (1-\mu_{\min})^{2N} \\
&\leqslant \mathbb{E}\left[\prod_{n=1}^{N}\left(1 - \left(1 + \exp\left(\frac{\beta\sqrt{2\ln(2d^{1/2}K)}}{\lambda}\right)\underbrace{\exp\left(-\beta\langle A^\star - a_n, \theta\rangle\right)}_{\blacktriangle}\right)^{-1}\right)\right] + \frac{1}{K} + (1-\mu_{\min})^{2N}.
\end{aligned} \tag{18}$$

Now, we define another event $\mathcal{E}_{(n)} := \{\langle A^\star - a_n, \theta\rangle \leqslant \Delta\}$. Based on $\mathcal{E}_{(n)}$ we analyze the $\blacktriangle$ term as follows.

$$\begin{aligned}
\exp\left(-\beta\langle A^\star - a_n, \theta\rangle\right) &= \mathbb{E}\big[\exp\left(-\beta\langle A^\star - a_n, \theta\rangle\right)\mathbb{I}_{\mathcal{E}_{(n)}}\big] + \mathbb{E}\big[\exp\left(-\beta\langle A^\star - a_n, \theta\rangle\right)\mathbb{I}_{\mathcal{E}_{(n)}^c}\big] \\
&\leqslant \exp(0)\,P(\mathcal{E}_{(n)}) + \exp\left(-\beta\Delta\right)P(\mathcal{E}_{(n)}^c) \\
&\leqslant P(\mathcal{E}_{(n)}) + (1 - P(\mathcal{E}_{(n)}))\exp\left(-\beta\Delta\right)
\end{aligned}$$

Plugging this back in Equation (18) we get,

$$P\left(A^\star \notin \mathcal{U}_{\mathcal{D}_0}\right) \leqslant \mathbb{E}\left[\prod_{n=1}^{N}\left(1-\left(1+\exp\left(\frac{\beta\sqrt{2\ln(2d^{1/2}K)}}{\lambda}\right)\left(\mathbb{I}_{\mathcal{E}_{(n)}}+(1-\mathbb{I}_{\mathcal{E}_{(n)}})\exp\left(-\beta\Delta\right)\right)\right)^{-1}\right)\right]+\frac{1}{K}+(1-\mu_{\min})^{2N} \tag{19}$$

Note that the random variable $\mathbb{I}_{\mathcal{E}_{(n)}}$ depends on the action sampling distribution μ. Denote the probability of sampling this action a_n by μ_n, and as before we have μ supported by $[\mu_{\min}, \mu_{\max}]$. We first analyze this for any arbitrary $n \in [N]$ and study the the distribution of $\mathbb{I}_{\mathcal{E}_{(n)}}$ conditionaled on $A^\star$. Without loss of generality, we first condition on $A^\star = \mathring{a}$ for some $\mathring{a} \in \mathcal{A}$. For that, let $\rho(\cdot)$ be the univariate Gaussian distribution and $\theta_a = \langle a, \theta\rangle$ for any action a.

$$\begin{aligned}
P\left(\mathbb{I}_{\mathcal{E}_{(n)}}=1 \,|\, A^\star=\mathring{a}\right) &= P\left(\mathbb{I}\left(\langle A^\star - a_n, \theta\rangle \leqslant \Delta\right)=1 \;\middle|\; A^\star=\mathring{a}\right)\\
&= \frac{1}{P\left(A^\star=\mathring{a}\right)} P\left(\mathbb{I}\left(\langle A^\star - a_n, \theta\rangle \leqslant \Delta\right)=1\,,\; A^\star=\mathring{a}\right)\\
&= \frac{1}{P\left(A^\star=\mathring{a}\right)} P\left(\mathbb{I}\left(\theta_{a_n} \geqslant \theta_{\mathring{a}} - \Delta\right)=1\,,\; \bigcap_{a\in\mathcal{A}}\{\theta_{\mathring{a}} \geqslant \theta_a\}\right)\\
&= \frac{1}{P\left(A^\star=\mathring{a}\right)} \int_{\mathbb{R}}\left[\int_{\theta_{\mathring{a}}-\Delta}^{\infty} d\rho(\theta)\right] d\rho(\theta_{\mathring{a}})\\
&= \frac{1}{P\left(A^\star=\mathring{a}\right)} \int_{\mathbb{R}}\left[\int_{\theta_{\mathring{a}}-\Delta}^{\theta_{\mathring{a}}} d\rho(\theta)\right] d\rho(\theta_{\mathring{a}}) \qquad\qquad (\text{since}\, \theta_a \leqslant \theta_{A^\star}\;\; \forall\; a)
\end{aligned} \tag{20}$$

Noticing that the term inside the integral can be represented as a distribution, we first find a normalizing constant to represent the probabilities. So, define

$$\Phi(\theta_{\mathring{a}}) = \int_{-\infty}^{\theta_{\mathring{a}}} (2\pi)^{-1/2}\exp(-x^2/2)\, dx \qquad ; \qquad g(\theta_{\mathring{a}}) = \frac{1}{\Phi(\theta_{\mathring{a}})}\int_{\theta_{\mathring{a}}-\Delta}^{\theta_{\mathring{a}}} d\rho(\theta)\,.$$

For fixed $\theta_{\mathring{a}}$, let $X_{\theta_{\mathring{a}}} \sim \text{Bernoulli}(1, g(\theta_{\mathring{a}}))$. With Eq. (20) and letting $d\mu(\theta_{\mathring{a}}) = \frac{\Phi(\theta_{\mathring{a}})}{P(A^\star=a_{\mathring{a}})} d\rho(\theta_{\mathring{a}})$ we have,

$$P\left(\mathbb{I}_{\mathcal{E}_{(n)}}=1 \,|\, A^\star=\mathring{a}\right) = \textstyle\int_{\mathbb{R}} P(X_{\theta_{\mathring{a}}}=1)\frac{\Phi(\theta_{\mathring{a}})}{P(A^\star=\mathring{a})} d\rho(\theta_{\mathring{a}}) = \int_{\mathbb{R}} P(X_{\theta_{\mathring{a}}}=1)\, d\mu(\theta_{\mathring{a}})\,. \tag{21}$$

Plugging this back in Equation (19) and upper bounding the probabilities we get,

$$\begin{aligned}
P\left(A^\star \notin \mathcal{U}_{\mathcal{D}_0}\right) &\leqslant \sum_{a\in\mathcal{A}}\int_{\mathbb{R}} P(X_{\theta_a}=0) d\mu(\theta_a) P\left(A^\star=a\right)\left(1-\left(1+\exp\left(\beta\left(\lambda^{-1}\sqrt{2\ln(2d^{1/2}K)}-\Delta\right)\right)\right)^{-1}\right)^N \cdot \mu_{\max}^N\\
&\quad + \frac{1}{K} + (1-\mu_{\min})^{2N}\\
&\leqslant \sum_{a\in\mathcal{A}}\int_{\mathbb{R}} P(X_{\theta_a}=0)\left(1-\left(1+\exp\left(\beta\left(\lambda^{-1}\sqrt{2\ln(2d^{1/2}K)}-\Delta\right)\right)\right)^{-1}\right)^N \cdot \mu_{\max}^N\, d\mu(\theta_a) P\left(A^\star=a\right)\\
&\quad + \frac{1}{K} + (1-\mu_{\min})^{2N}\\
&\leqslant \int_{\mathbb{R}} \mathop{\mathbb{E}}_{\mathring{a}\in\mathcal{A}}\left[\left(1-\left(1+\exp\left(\beta\left(\lambda^{-1}\sqrt{2\ln(2d^{1/2}K)}-(1-X_{\theta_{\mathring{a}}})\Delta\right)\right)\right)^{-1}\right)^N\right] \cdot \mu_{\max}^N d\mu(\theta_{\mathring{a}}) + \frac{1}{K} + (1-\mu_{\min})^{2N}\,,
\end{aligned} \tag{22}$$

where $\mu_{\max}$ is used to obtain the exponent N by accounting for the sampling distribution μ, and last step follows from the uniformity of each action being optimal. Finally, we need to find the supremum of $g(\theta_{\mathring{a}})$ and hence Equation (22). Recall that,

$$g(\theta_{\mathring{a}}) = \frac{1}{\Phi(\theta_{\mathring{a}})}\int_{\theta_{\mathring{a}}-\Delta}^{\theta_{\mathring{a}}} d\rho(\theta) = \frac{\int_{\theta_{\mathring{a}}-\Delta}^{\theta_{\mathring{a}}} d\rho(\theta)}{\int_{-\infty}^{\theta_{\mathring{a}}} d\rho(\theta)} = \frac{\int_{-\infty}^{\theta_{\mathring{a}}} d\rho(\theta) - \int_{-\infty}^{\theta_{\mathring{a}}-\Delta} d\rho(\theta)}{\int_{-\infty}^{\theta_{\mathring{a}}} d\rho(\theta)} = 1 - h_{\Delta}(\mathring{a})$$

, where $h_{\Delta}(\mathring{a}) := \frac{\int_{-\infty}^{\theta_{\mathring{a}}-\Delta} d\rho(\theta)}{\int_{-\infty}^{\theta_{\mathring{a}}} d\rho(\theta)}$. Setting $\nabla_{\mathring{a}} h_{\Delta}(\mathring{a}) = 0$ and analyzing $\nabla^2_{\mathring{a}} h_{\Delta}(\mathring{a}) > 0$, we find that

$$g(\theta_{\mathring{a}}) \leqslant 1 - \Delta \exp\left(-\frac{(2\theta_{\mathring{a}} - \Delta)\Delta}{2}\right) \leqslant \min(1, \Delta)\ . \tag{23}$$

Finally we, decompose Equation (22) based on the event $\mathcal{E}_2 := \{X_{\theta_{\mathring{a}}} = 0\}$, and upper bound the probability to simplify. Setting $\Delta = \ln(K\beta)/\beta$, we conclude with the following bound:

$$P\left(A^{\star} \notin \mathcal{U}_{\mathcal{D}_0}\right) \leqslant \left(1 - \left(1 + \exp\left(\beta\left(\lambda^{-1}\sqrt{2\ln(2d^{1/2}K)} - (A-1)\min(1, \ln(K\beta)/\beta)\right)\right)\right)^{-1}\right)^{N} + \frac{1}{K} + (1-\mu_{\min})^{2N} \tag{24}$$

□

Theorem A.8. *For any confidence $\delta_1 \in (0, \frac{1}{3})$ and offline preference dataset size $N > 2$, the simple Bayesian regret of the learner Υ is upper bounded with probability of at least $1 - 3\delta_1$ by,*

$$\begin{aligned}
\mathfrak{SR}_K^{\Upsilon}(\pi^{\star}_{K+1}, \pi^{\star}) &\leqslant \sqrt{\frac{20\delta_2 A \ln\left(\frac{2KA}{\delta_1}\right)}{2K\left(1 + \ln\frac{A}{\delta_1}\right) - \ln\frac{A}{\delta_1}}} \quad \textit{with,} \\
\gamma_{\beta,\lambda,N} &= \exp\left(-\beta B\sqrt{2\ln(2d^{1/2}N)}/\lambda - \beta\Delta_{\min}\right) + \frac{1}{N} \ \textit{, and} \\
\delta_2 &= 2\exp\left(-N\left(1+\gamma_{\beta,\lambda,N}\right)^2\right) + \exp\left(-\frac{N}{4}(1-\gamma_{\beta,\lambda,N})^3\right).
\end{aligned} \tag{25}$$

For a fixed $N > 2$, and large A, the simple regret bound is $\widetilde{\mathcal{O}}\left(\sqrt{AK^{-1}}\right)$. Note that this bound converges to zero exponentially fast as $N \to \infty$ and as the rater tends to an expert (large β, λ). In addition, as the number of online episodes K gets large, PSPL is able to identify the best policy (arm) with probability at least $(1 - 3\delta_1)$.

A.3 Constructing Surrogate Loss Function

Lemma A.9. *At episode k, the MAP estimate of $(\theta, \vartheta, \eta)$ can be constructed by solving the following equivalent optimization problem:*

$$\begin{aligned}
(\theta_{opt}, \vartheta_{opt}, \eta_{opt}) &= \operatorname*{argmax}_{\theta,\vartheta,\eta} \Pr(\theta, \vartheta, \eta \,|\, \mathcal{D}_k) \\
&\equiv \operatorname*{argmin}_{\theta,\vartheta,\eta} \mathcal{L}_1(\theta,\vartheta,\eta) + \mathcal{L}_2(\theta,\vartheta,\eta) + \mathcal{L}_3(\theta,\vartheta,\eta)\ \textit{, where,} \\
\mathcal{L}_1(\theta,\vartheta,\eta) &:= -\sum_{t=1}^{k-1}\left[\beta\langle\tau_t^{(Y_t)}, \vartheta\rangle - \ln\left(e^{\beta\langle\tau_t^{(0)},\vartheta\rangle} + e^{\beta\langle\tau_t^{(1)},\vartheta\rangle}\right)\right. \\
&\qquad \left. + \sum_{j=0}^{1}\sum_{h=1}^{H-1}\ln\Pr_{\eta}\left(s_{t,h+1}^{(j)} \,|\, s_{t,h}^{(j)}, a_{t,h}^{(j)}\right)\right], \\
\mathcal{L}_2(\theta,\vartheta,\eta) &:= -\sum_{n=1}^{N}\left[\beta\langle\overline{\tau}_n^{(\overline{Y}_n)}, \vartheta\rangle - \ln\left(e^{\beta\langle\overline{\tau}_n^{(0)},\vartheta\rangle} + e^{\beta\langle\overline{\tau}_n^{(1)},\vartheta\rangle}\right)\right], \\
\mathcal{L}_3(\theta,\vartheta,\eta) &:= \frac{\lambda^2}{2}\|\theta - \vartheta\|_2^2 - SA\sum_{i=1}^{S}(\boldsymbol{\alpha}_{0,i} - 1)\ln\eta_i \\
&\qquad + \frac{1}{2}(\theta - \mu_0)^T\Sigma_0^{-1}(\theta - \mu_0).
\end{aligned} \tag{26}$$

Proof. We first analyze the posterior distribution of ϑ, θ, η given the dataset $\mathcal{D}_k$ at the beginning of episode k, and then optimize it by treating these random variables as parameters.

$$\begin{aligned}
\underset{\theta,\vartheta,\eta}{\operatorname{argmax}}\ \Pr(\theta,\vartheta,\eta \,|\, \mathcal{D}_k) &= \underset{\theta,\vartheta,\eta}{\operatorname{argmax}}\ \Pr(\mathcal{D}_k \,|\, \theta,\vartheta,\eta)\cdot\Pr(\theta,\vartheta,\eta) \\
&= \underset{\theta,\vartheta,\eta}{\operatorname{argmax}}\ \ln\Pr(\mathcal{D}_k \,|\, \theta,\vartheta,\eta) + \ln\Pr(\theta,\vartheta,\eta) \\
&= \underset{\theta,\vartheta,\eta}{\operatorname{argmax}} \underbrace{\ln\Pr(\mathcal{H}_k \,|\, \mathcal{D}_0,\theta,\vartheta,\eta)}_{\mathcal{L}_1} + \underbrace{\ln\Pr(\mathcal{D}_0 \,|\, \theta,\vartheta,\eta)}_{\mathcal{L}_2} + \underbrace{\ln\Pr(\theta,\vartheta,\eta)}_{\mathcal{L}_3}
\end{aligned} \tag{27}$$

Then,

$$\begin{aligned}
\mathcal{L}_1 &= \sum_{t=1}^{k-1} \ln\Pr\left(\left(\tau_t^{(0)},\tau_t^{(1)},Y_t\right) \,\middle|\, \mathcal{D}_t,\theta,\vartheta,\eta\right) \\
&= \sum_{t=1}^{k-1} \ln\Pr\left(Y_t \,|\, \tau_t^{(0)},\tau_t^{(1)},\theta,\vartheta,\eta\right) + \ln\Pr\left(\tau_t^{(0)},\tau_t^{(1)} \mid \mathcal{D}_t,\theta,\vartheta,\eta\right) \\
&= \sum_{t=1}^{k-1}\left[\beta\langle\tau_t^{(Y_t)},\vartheta\rangle - \ln\left(e^{\beta\langle\tau_t^{(0)},\vartheta\rangle} + e^{\beta\langle\tau_t^{(1)},\vartheta\rangle}\right) + \sum_{j=0}^{1}\sum_{h=1}^{H-1}\ln\Pr_\eta\left(s_{t,h+1}^{(j)} \,|\, s_{t,h}^{(j)},a_{t,h}^{(j)}\right)\right] \\
\mathcal{L}_2 &= \sum_{n=1}^{N} \ln\Pr\left(\left(\overline{\tau}_n^{(0)},\overline{\tau}_n^{(1)},\overline{Y}_n\right) \,\middle|\, \theta,\vartheta,\eta\right) \\
&= \sum_{n=1}^{N} \ln\Pr\left(\overline{Y}_n \,|\, \overline{\tau}_n^{(0)},\overline{\tau}_n^{(1)},\theta,\vartheta,\eta\right) + \underbrace{\ln\Pr\left(\overline{\tau}_n^{(0)},\overline{\tau}_n^{(1)} \mid \theta,\vartheta,\eta\right)}_{\text{indep. of } \theta,\vartheta,\eta \ \Rightarrow\ \text{constant}} \\
&= \sum_{n=1}^{N} \beta\langle\overline{\tau}_n^{(\overline{Y}_n)},\vartheta\rangle - \ln\left(e^{\beta\langle\overline{\tau}_n^{(0)},\vartheta\rangle} + e^{\beta\langle\overline{\tau}_n^{(1)},\vartheta\rangle}\right) + \text{constant} \\
\mathcal{L}_3 &= \ln\Pr(\vartheta \,|\, \theta) + \ln\Pr(\theta) + \ln\Pr(\eta) \\
&= \frac{d}{2}\ln\left(\frac{2\pi}{\lambda^2}\right) - \frac{\lambda^2}{2}\|\theta-\vartheta\|_2^2 - \frac{1}{2}\ln\left(|2\pi\Sigma_0|\right) - \frac{1}{2}(\theta-\mu_0)^T\Sigma_0^{-1}(\theta-\mu_0) + SA\sum_{i=1}^{S}(\boldsymbol{\alpha}_{0,i}-1)\ln\eta_i.
\end{aligned} \tag{28}$$

Hence, final surrogate loss function is

$$\begin{aligned}
\mathcal{L}(\theta,\vartheta,\eta) &= \mathcal{L}_1(\theta,\vartheta,\eta) + \mathcal{L}_2(\theta,\vartheta,\eta) + \mathcal{L}_3(\theta,\vartheta,\eta), \qquad \text{where} \\
\mathcal{L}_1(\theta,\vartheta,\eta) &= -\sum_{t=1}^{k-1}\left[\beta\langle\tau_t^{(Y_t)},\vartheta\rangle - \ln\left(e^{\beta\langle\tau_t^{(0)},\vartheta\rangle} + e^{\beta\langle\tau_t^{(1)},\vartheta\rangle}\right) + \sum_{j=0}^{1}\sum_{h=1}^{H-1}\ln\Pr_\eta\left(s_{t,h+1}^{(j)} \,|\, s_{t,h}^{(j)},a_{t,h}^{(j)}\right)\right], \\
\mathcal{L}_2(\theta,\vartheta,\eta) &= -\sum_{n=1}^{N}\left[\beta\langle\overline{\tau}_n^{(\overline{Y}_n)},\vartheta\rangle - \ln\left(e^{\beta\langle\overline{\tau}_n^{(0)},\vartheta\rangle} + e^{\beta\langle\overline{\tau}_n^{(1)},\vartheta\rangle}\right)\right], \\
\mathcal{L}_3(\theta,\vartheta,\eta) &= \frac{\lambda^2}{2}\|\theta-\vartheta\|_2^2 + \frac{1}{2}(\theta-\mu_0)^T\Sigma_0^{-1}(\theta-\mu_0) - SA\sum_{i=1}^{S}(\boldsymbol{\alpha}_{0,i}-1)\ln\eta_i.
\end{aligned} \tag{29}$$

Finally the problem becomes equivalent as follows:

$$(\theta_{opt},\vartheta_{opt},\eta_{opt}) = \underset{\theta,\vartheta,\eta}{\operatorname{argmax}}\ \Pr(\theta,\vartheta,\eta \,|\, \mathcal{D}_k) \equiv \underset{\theta,\vartheta,\eta}{\operatorname{argmin}}\ \mathcal{L}(\theta,\vartheta,\eta) \tag{30}$$

□

A.4 Practical PSPL (cont.)

A.4.1 Estimating Rater Competence in Practice

There are two main methods of estimating rater competence in practice:

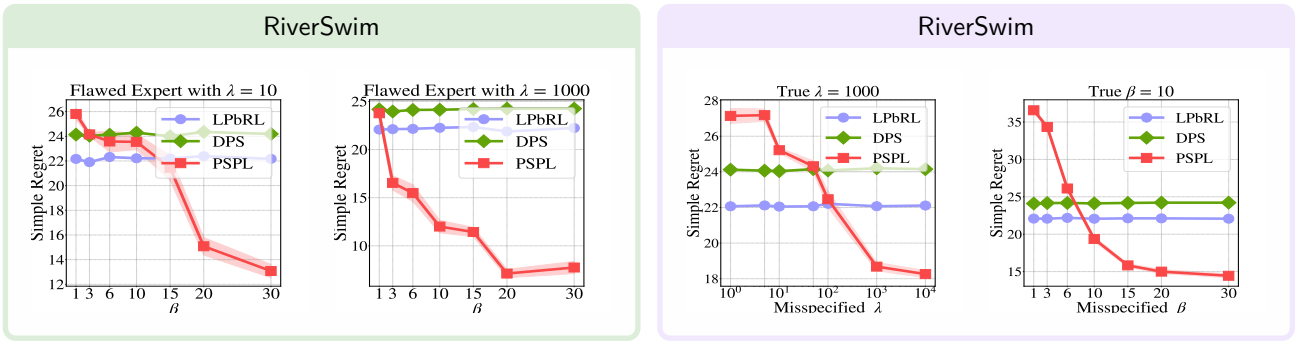


Figure 5: Sensitivity to flawed expert policy with $\lambda = \{10, 10^3\}$, and misspecified competence.

1. Based on maximum likelihood estimation (MLE). Similar idea has been proposed to estimate the expertise level in imitation learning Beliaev et al. (2022); Beliaev and Pedarsani (2025).

2. The second method is to simply look at the entropy of the empirical distribution of the action in the offline dataset. Suppose the empirical distribution of ζ. Then we use $c/\mathcal{H}(\zeta)$ as an estimation for β, where $c > 0$ is a hyperparameter. The intuition is that for smaller β, the net state-action pair visit counts tend to be more uniform and thus the entropy will be larger. This is an unsupervised approach and agnostic to specific offline data generation processes. The knowledgeability λ is not quite 'estimable' because for a single environment, even though we know the true environment θ and the expert's knowledge ϑ, we only have one pair of observations generated with the same ϑ. Thus, the variance of the estimation for λ could be infinite. However, exact estimation of λ is not necessary as we show that the algorithm is robust to misspecified λ through experiments in Section 6.

A.4.2 Ablation Studies

The Bootstrapped PSPL algorithm in Section 5 requires a knowledge of rater's parameters λ, β. We study the sensitivity of the algorithm's performance to mis-specification of these parameters.

(i) **Different Preference Generation Expert Policy.** Though the learning agent assumes Equation (2) as the expert's generative model, we consider it to use a deterministic greedy policy. Trajectories $\overline{\tau}_n^{(0)}$ and $\overline{\tau}_n^{(1)}$ are sampled, and then choose $\overline{Y}_n = \text{argmax}_{i\in\{0,1\}} \beta\langle\overline{\tau}_n^{(i)}, \vartheta\rangle$, where $\vartheta \sim \mathcal{N}(\theta, \mathbf{I}_d/\lambda^2)$. In Figure 5, we see that even when the learning agent's assumption of the rater policy is *flawed*, PSPL significantly outperforms the baselines. (ii) **Misspecified Competence parameters.** First, we generate offline data with true $\lambda = 10^3$ but PSPL uses a misspecified λ. Second, we generate offline data with true $\beta = 10$ but PSPL uses a misspecified β. Figure 5 shows that although the performance of PSPL decreases as the degree of flawness increases, it still outperforms the baselines.

A.4.3 Experiments on Image Generation Tasks (cont.)

We instantiate our framework on the Pick-a-Pic dataset of human preferences for text–to–image generation Kirstain et al. (2023). Overall, the dataset contains over 500,000 examples and 35,000 distinct prompts. Each example contains a prompt, sequence generations of two images, and a label for which image is preferred. We let each generation be a trajectory, so the dataset contains trajectory preferences $\mathcal{D}_0 = (\tau_i^+, \tau_i^-, y_i)_{i=1}^N$ with $y_i = 1$ iff $\tau_i^+ \succ \tau_i^-$. Each trajectory $\tau = (p, z_{0:T})$ is the entire latent denoising chain of length T for prompt p sampled from some prompt distribution.

Following Black et al. (2024); Zhang et al. (2024b), text–to–image generation is a finite–horizon MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P)$: $s_t = (p, z_t), a_t = \epsilon_t, z_{t+1} = f_\rho(z_t, \epsilon_t)$, where $z_t \in \mathbb{R}^d$ is the latent, a_t is the noise ϵ_t predicted by the policy $\pi_\theta(a_t \mid s_t)$, and $f_{\boldsymbol{\alpha}}$ is the deterministic DDPM transition with frozen scheduler ρ. Please see Black et al. (2024) for a comprehensive and detailed discussion on modeling diffusion as a MDP; we follow the same approach. The episode horizon is $H = T = 50$. For each trajectory, we adopt the additive embedding as discussed in the paper: $\phi(\tau) = \sum_{t=1}^T \phi(s_t, a_t)$ with $\phi(s_t, a_t) = \big[\text{CLIP}(p, \text{Dec}(z_T)), \|\epsilon_t\|_2, t/T\big]$, where $\text{Dec}(\cdot)$ is the standard VAE decoder shipped with Stable-Diffusion Rombach et al. (2022). As in the paper, the model assumes that rater of competence λ, β follows the Bradley-Terry model i.e. $\Pr(Y = 1 \mid \tau^+, \tau^-, \vartheta, \beta) = \sigma\big(\beta\langle\vartheta, \phi(\tau^+) - \phi(\tau^-)\rangle\big)$, where $\sigma(\cdot)$ is the sigmoid link function.

To model the reward parameters, we let the uninformed prior be $\nu_0 = \mathcal{N}(\mu_0, \Sigma_0)$, $\mu_0 = 0$, and diagonal $\Sigma_0 = \text{diag}(\sigma^2_{\text{clip}}, \sigma^2_{\text{noise}}, \sigma^2_{\text{time}})$, where $\sigma^2_{\text{clip}}, \sigma^2_{\text{noise}}$ are empirical variances from 10k random chains and $\sigma^2_{\text{time}} = 1/12$. Regarding modeling of transitions, the physical scheduler f_ρ is known; uncertainty remains only in the ℓ_2-norm of ϵ_t. We discretize this norm into C=10 bins as b_i for $i \in [C]$ and model $P(b_i \mid s_t) \sim \text{Dir}(\boldsymbol{\alpha}_0)$, $\boldsymbol{\alpha}_0 = \mathbf{1}_C$. The resulting Dirichlet counts are updated from both $\mathcal{D}_0$ and online episodes.

Since we do not know the optimal generation sequence, minimizing simple regret is equivalent to maximizing expected reward in the final online episode. To evaluate this final generation, we conduct $L = 10$ rollouts and compute the average reward with a weighted ensemble of automatic quality metrics, similar to Clark et al. (2023); Xu et al. (2023) i.e. $r_\theta(\cdot) = 0.7 * \text{ImageReward–v2} + 0.3 * \text{Aesthetic–LAION}$ where ImageReward–v2 is a reward model from Xu et al. (2023), and Aesthetic–LAION is a reward model hosted on HuggingFace eV (2025). For evaluation, we sweep $N_{\text{off}} = 50$k preference triplets for the prior, reserve $N_{\text{val}} = 15$k examples for evaluation, and create an exploration pool of 10k unseen examples. For implementation, we use a 2 layer MLP, 512 GELU units, and outputs $(\mu_\theta)_t$ and $\log((\sigma_\theta)_t)$, where $(\mu_\theta)_t$ is the predicted noise vector the agent believes will best denoise z_t and $(\sigma_\theta)_t$ controls exploratory perturbations around that prediction, enabling posterior sampling for PSPL. This is analogous to solving the unconstrained optimization Problem (9) in the function approximation setting.

For the online phase, we use the above reward model $r_\theta(\cdot)$ for the BT preference model (see Equation (2)), with an expert rater (i.e. $\lambda, \beta \to \infty$), similar to Kirstain et al. (2023). Since, we show that PSPL is robust to mis-specifications in rater competence (please see Appendix A.4.2), we use an expert rater for ease of comparison.

Training details. Following Zhang et al. (2024b), we first cluster the prompts in the dataset to obtain a mapping from $\text{cluster}_j \to (\mathcal{D}_0)_j$, where $j \in [J]$ is the cluster index out of J clusters, and $(\mathcal{D}_0)_j$ is the dataset of trajectories and corresponding preference labels for prompts in prompt cluster_j i.e. $(\mathcal{D}_0)_j = (\tau^+_{j,i}, \tau^-_{j,i}, y_{j,i})_{i=1}^{N_j}$, where $\tau^+_{j,\cdot}$ and $\tau^-_{j,\cdot}$ are the *winning* and *losing* trajectory generations given a prompt from cluster_j for all $j \in [J]$. For tractability, we compress the training images to be 128×128 pixels, and optimize for $K = 100k$ episodes for each cluster $j \in [J]$. Future direction of this work will consider incorporating prompt information as a prior to the MLP, resulting in prompt conditioned inference. However, that is beyond the current scope of the paper. Finally, all experiments are run on NVIDIA GeForce RTX 5080, GPU 16GB, and Memory DDR5 64 GB. Training times for all algorithms are given in Table A.4.3 and comprehensive validation results are shown in Figure 6.

Table 1: Average training times of baselines over 5 independent runs.

	DPS	PSPL
Training Time (h)	3.52 ± 0.11	3.73 ± 0.09

B Auxiliary Definitions and Lemmas

Lemma B.1. *Let X be the sum of L i.i.d. Bernoulli random variables with mean $p \in (0,1)$. Let $q \in (0,1)$, then*

$$\Pr(X \leqslant qL) \leqslant \exp\left(-2L(q-p)^2\right), \qquad \text{if } q < p,$$
$$\Pr(X \geqslant qL) \leqslant \exp\left(-2L(q-p)^2\right), \qquad \text{if } q > p.$$

Proof. Both inequalities can be obtained by applying Hoeffding's Inequality (see B.9). □

Definition B.2. **α-dependence in Russo and Van Roy (2013).** For $\alpha > 0$ and function class $\mathcal{Z}$ whose elements are with domain $\mathcal{X}$, an element $x \in \mathcal{X}$ is α-dependent on the set $\mathcal{X}_n := \{x_1, x_2, \cdots, x_n\} \subset \mathcal{X}$ with respect to $\mathcal{Z}$, if any pair of functions $z, z' \in \mathcal{Z}$ with $\sqrt{\sum_{i=1}^n \left(z(x_i) - z'(x_i)\right)^2} \leqslant \alpha$ satisfies $z(x) - z'(x) \leqslant \alpha$. Otherwise, x is α-independent on $\mathcal{X}_n$ if it does not satisfy the condition.

Definition B.3. Eluder dimension in Russo and Van Roy (2013). For $\alpha > 0$ and function class $\mathcal{Z}$ whose elements are with domain $\mathcal{X}$, the Eluder dimension $\dim_E(\mathcal{Z}, \alpha)$, is defined as the length of the longest possible sequence of elements in $\mathcal{X}$ such that for some $\alpha' \geqslant \alpha$, every element is α' independent of its predecessors.

Definition B.4. Covering number. Given two functions l and u, the bracket $[l, u]$ is the set of all functions f satisfying $l \leqslant f \leqslant u$. An α-bracket is a bracket $[l, u]$ with $\|u - l\| < \alpha$. The covering number $N_{[\cdot]}(\mathcal{F}, \alpha, \|\cdot\|)$ is the minimum number of α-brackets needed to cover $\mathcal{F}$.

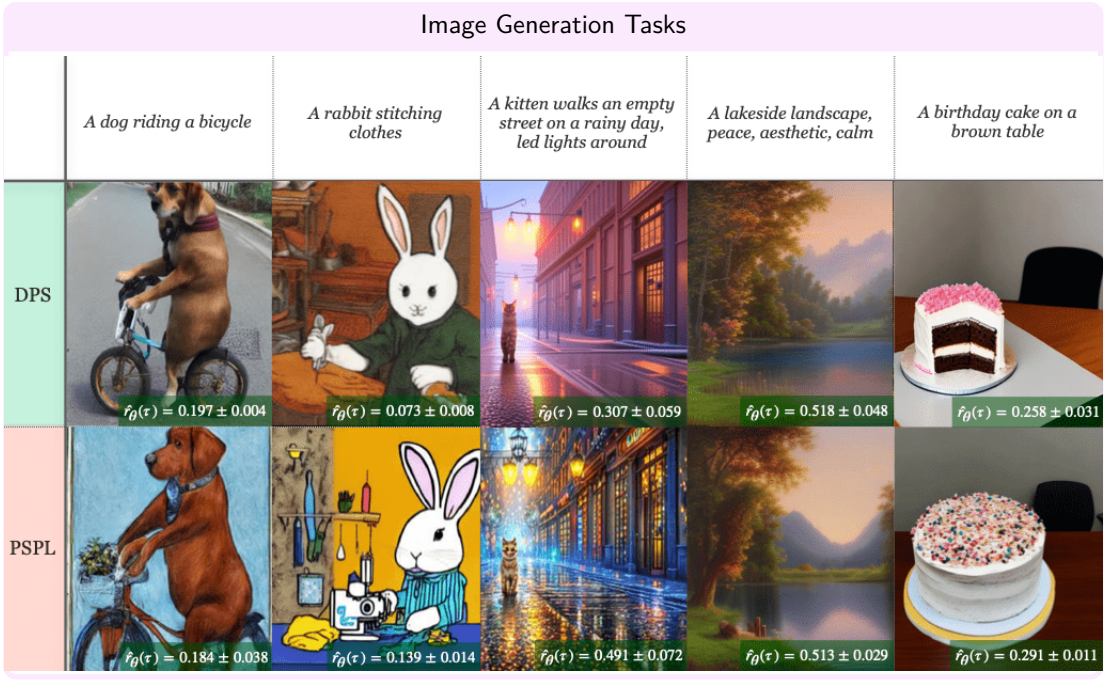


Figure 6: Sample image generations with final image reward $\widehat{r}_\theta(\cdot)$ over 5 independent runs. Images are enlarged for clarity.

Lemma B.5. ***(Linear Preference Models Eluder dimension and Covering number).*** *For the case of d-dimensional generalized trajectory linear feature models $r_\xi(\xi_H) := \langle \phi(\xi_H), \mathbf{w}_r \rangle$, where $\phi : Traj \to \mathbb{R}^{dim_{\mathbb{T}}}$ is a known $dim_{\mathbb{T}}$ dimension feature map satisfying $\|\psi(\xi_H)\|_2 \leqslant B$ and $\theta \in \mathbb{R}^d$ is an unknown parameter with $\|\mathbf{w}_r\|_2 \leqslant \kappa_w$. Then the α-Eluder dimension of $r_\xi(\xi_H)$ is at most $\mathcal{O}(dim_{\mathbb{T}} \log(B\kappa_w/\alpha))$. The α - covering number is upper bounded by $\left(\frac{1+2B\kappa_w}{\alpha}\right)^{dim_{\mathbb{T}}}$.*

Let $(X_p, Y_p)_{p=1,2,\ldots}$ be a sequence of random elements, $X_p \in X$ for some measurable set X and $Y_p \in \mathbb{R}$. Let $\mathcal{F}$ be a subset of the set of real-valued measurable functions with domain X. Let $\mathbb{F} = (\mathbb{F}_p)_{p=0,1,\cdots}$ be a filtration such that for all $p \geqslant 1, (X_1, Y_1, \cdots, X_{p-1}, Y_{p-1}, X_p)$ is $\mathbb{F}_{p-1}$ measurable and such that there exists some function $f_\star \in \mathcal{F}$ such that $\mathbb{E}[Y_p \mid \mathbb{F}_{p-1}] = f_*(X_p)$ holds for all $p \geqslant 1$. The (nonlinear) least square predictor given $(X_1, Y_1, \cdots, X_t, Y_t)$ is $\widehat{f}_t = \operatorname{argmin}_{f \in \mathcal{F}} \sum_{p=1}^t (f(X_p) - Y_p)^2$. We say that Z is conditionally κ-subgaussion given the σ-algebra $\mathbb{F}$ is for all $\lambda \in \mathbb{R}, \log \mathbb{E}[\exp(\lambda Z) \mid \mathbb{F}] \leqslant \frac{1}{2}\lambda^2\kappa^2$. For $\alpha > 0$, let N_α be the $\|\cdot\|_\infty$-covering number of $\mathcal{F}$ at scale α. For $\beta > 0$, define

$$\mathcal{F}_t(\beta) = \left\{ f \in \mathcal{F} : \sum_{p=1}^t \left(f(X_p) - \widehat{f}_t(X_p) \right)^2 \leqslant \beta \right\}. \tag{31}$$

Lemma B.6. *(Theorem 5 of Ayoub et al. (2020)). Let $\mathbb{F}$ be the filtration defined above and assume that the functions in $\mathcal{F}$ are bounded by the positive constant $C > 0$. Assume that for each $s \geqslant 1, (Y_p - f_*(X_p))$ is conditionally σ-subgaussian given $\mathbb{F}_{p-1}$. Then, for any $\alpha > 0$, with probability $1-\delta$, for all $t \geqslant 1, f_* \in \mathcal{F}_t(\beta_t(\delta, \alpha))$, where*

$$\beta_t(\delta, \alpha) = 8\sigma^2 \log(2N_\alpha/\delta) + 4t\alpha \left(C + \sqrt{\sigma^2 \log(4t(t+1)/\delta)} \right).$$

Lemma B.7. *(Lemma 5 of Russo and Van Roy (2013)). Let $\mathcal{F} \in B_\infty(X, C)$ be a set of functions bounded by $C > 0$, $(\mathcal{F}_t)_{t \geqslant 1}$ and $(x_t)_{t \geqslant 1}$ be sequences such that $\mathcal{F}_t \subset \mathcal{F}$ and $x_t \in \mathcal{X}$ hold for $t \geqslant 1$. Let $\mathcal{F}|_{x_{1:t}} = \{(f(x_1), \ldots, f(x_t)) : f \in \mathcal{F}\} (\subset \mathbb{R}^t)$ and for $S \subset \mathbb{R}^t$, let $\operatorname{diam}(S) = \sup_{u,v \in S} \|u - v\|_2$ be the diameter of S. Then, for any $T \geqslant 1$ and $\alpha > 0$ it, holds that*

$$\sum_{t=1}^T \operatorname{diam}\left(\mathcal{F}_t|_{x_t} \right) \leqslant \alpha + C(d \wedge T) + 2\delta_T \sqrt{dT},$$

where $\delta_T = \max_{1 \leqslant t \leqslant T} \operatorname{diam}\left(\mathcal{F}_t|_{x_{1:t}} \right)$ and $d = \dim_{\mathcal{E}}(\mathcal{F}, \alpha)$.

Lemma B.8. *If* $(\beta_t \geqslant 0 \mid t \in \mathbb{N})$ *is a nondecreasing sequence and* $\mathcal{F}_t := \left\{ f \in \mathcal{F} : \left\| f - \widehat{f}_t^{LS} \right\|_{2,E_t} \leqslant \sqrt{\beta_t} \right\}$*, where* $\widehat{f}_t^{LS} \in \arg\min_{f \in \mathcal{F}} L_{2,t}(f)$ *and* $L_{2,t}(f) = \sum_1^{t-1} \left(f\left(A_t\right) - R_t \right)^2$*, then for all* $T \in \mathbb{N}$ *and* $\epsilon > 0$*,*

$$\sum_{t=1}^{T} \mathbf{1}\left(w_{\mathcal{F}_t}\left(A_t\right) > \epsilon \right) \leqslant \left(\frac{4\beta_T}{\epsilon^2} + 1 \right) \dim_E(\mathcal{F}, \epsilon)$$

where $w_{\mathcal{F}}(a) := \sup_{f \in \mathcal{F}} f(a) - \inf_{f \in \mathcal{F}} f(a)$ *denotes confidence interval widths.*

Theorem B.9. ***Hoeffding's inequality****(Hoeffding, 1994). Let* $X_1, X_2, \ldots, X_n$ *be independent random variables that are sub-Gaussian with parameter* σ*. Define* $S_n = \sum_{i=1}^n X_i$*. Then, for any* $t > 0$*, Hoeffding's inequality provides an upper bound on the tail probabilities of* S_n*, which is given by:*

$$\Pr\left(|S_n - \mathbb{E}[S_n]| \geqslant t \right) \leqslant 2\exp\left(-\frac{t^2}{2n\sigma^2} \right).$$

This result emphasizes the robustness of the sum S_n *against deviations from its expected value, particularly useful in applications requiring high confidence in estimations from independent sub-Gaussian observations.*

Lemma B.10. *(Lemma F.4. in Dann et al. (2017)) Let* $\mathcal{F}_i$ *for* $i = 1 \ldots$ *be a filtration and* $X_1, \ldots X_n$ *be a sequence of Bernoulli random variables with* $\mathbb{P}\left(X_i = 1 \mid \mathcal{F}_{i-1}\right) = P_i$ *with* P_i *being* $\mathcal{F}_{i-1}$*-measurable and* X_i *being* $\mathcal{F}_i$ *measurable. It holds that*

$$\mathbb{P}\left(\exists n : \sum_{t=1}^{n} X_t < \sum_{t=1}^{n} P_t/2 - W \right) \leqslant e^{-W}$$