Fine-tuning Behavioral Cloning Policies with Preference-Based Reinforcement Learning

Maël Macuglia[1], **Paul Friedrich**[1,2], **Giorgia Ramponi**[1,2]
mael.macuglia@icloud.com, {paul.friedrich, giorgia.ramponi}@uzh.ch

[1]**Department of Informatics, University of Zurich, Switzerland**
[2]**ETH AI Center, Zurich, Switzerland**

Abstract

Deploying reinforcement-learning (RL) controllers in robotics, industry, and health care is blocked by two coupled obstacles: reward misspecification (informal goals are hard to encode as a safe numeric signal) and data-hungry exploration. We tackle these issues with a two-stage framework that begins from a reward-free dataset of expert demonstrations and refines the policy online using preference-based human feedback. We give the first principled analysis of this two-stage paradigm. In our work, we formulate a unified algorithm that (i) clones demonstrations offline to obtain a safe warm-start policy and (ii) fine-tunes it online with preference-based RL, integrating the two signals through an uncertainty-weighted objective. Then, we derive regret bounds that shrink with the demonstration counts and reflect reduced uncertainty.

1 Introduction

Deploying reinforcement-learning (RL) (Sutton & Barto, 2018) systems on physical robots, industrial processes, and health-care problems remains notoriously difficult for two intertwined reasons. First, reward misspecification: even experienced domain experts often find it hard to translate informal task goals into a numeric signal that is simultaneously accurate and safe (Leike et al., 2018). Second, exploration is both risky and data-hungry (Dulac-Arnold et al., 2019): a policy that begins from scratch can damage hardware, or user trust, long before it gathers enough experience to learn anything useful. Recent applied works (Nair et al., 2020; Kostrikov et al., 2022; Tang et al., 2025; Park et al., 2024; Tirinzoni et al., 2025) alleviate these issues by pre-training policies offline and fine-tuning them with online RL, but such methods assume direct access to (or observation of) the true reward, an assumption that rarely holds in practice. A more realistic strategy is to start with a reward-free corpus of expert behavior and allow the experts to refine that behavior online. The refinement signal can take several forms: explicit numeric rewards from an operator, pairwise comparisons of trajectories, or scalar ratings that train a reward model for RL from human feedback (RLHF). Variants of this idea already power modern dialogue agents such as ChatGPT, which first imitate curated demonstrations of desirable responses and are then fine-tuned via RLHF on preference-derived rewards (Ouyang et al., 2022). Similar approaches reach near-expert scores in Atari and MuJoCo by combining brief expert play with thousands of comparison queries (Christiano et al., 2017), or recover usable rewards for real-robot manipulation by ranking tele-operated clips before on-hardware fine-tuning (Brown et al., 2020).

This paper formalizes a principled two-stage procedure: offline demonstrations supply a safe warm start, while lightweight online feedback repairs the blind spots of the behavioral-cloning policy. Merging these two information sources yields both higher sample efficiency and stronger safety guarantees. Demonstrations guide the agent away from dangerous or hard-to-reach regions of state space, so exploration seldom visits unsafe states. Theoretically, when the offline data already explain many state–action pairs, regret bounds shrink because the set of plausible optimal policies is greatly reduced. Pragmatically, preference queries remain easy to elicit even when explicit rewards are not, which lets non-technical stakeholders participate in the corrective loop. We conclude by detailing our contributions:

- We develop a novel policy confidence set framework based on Hellinger distances between trajectory distributions. By separating the policy and transition components of the MLE objective, we extend Foster et al. (2024)'s framework to obtain distribution-level guarantees. This confidence set (Theorem 4) is geometrically interpretable as a Hellinger ball in trajectory distribution space while providing a corresponding constraint on allowable policies. Its radius decreases with the offline sample size, effectively leveraging demonstration data to restrict the policy search space. The approach generalizes to various policy and transition model classes through appropriate covering number arguments.
- We adapt the online preference-based learning framework to leverage our offline estimation components, resulting in BRIDGE (Algorithm 1). By constraining policy comparisons to our confidence set and initializing with the MLE transition estimate, our analysis yields a regret bound (Theorem 8) that exhibits optimal $\sqrt{T}$ dependence while demonstrating how offline data reduces online regret. The bound contains terms that explicitly diminish with increasing offline sample size n, and importantly, for any fixed horizon T, as $n \to \infty$, the regret approaches zero. This theoretical result confirms that high-quality offline demonstrations can dramatically improve online learning efficiency, creating a principled bridge between imitation learning and preference-based fine-tuning.

2 Related work

Behavioral Cloning (BC). BC reformulates RL as supervised learning on expert (state, action) pairs, pioneered by road-following systems like ALVINN (Pomerleau, 1988). Recent theoretical advances by Foster et al. (2024) establish horizon-free sample complexity bounds under deterministic policies and sparse rewards. Other algorithms such as DAgger (Ross et al., 2011) mitigate the covariate shift during deployment through iterative expert corrections, achieving no-regret guarantees (Ross et al., 2011). Our method inherits BC's simplicity but circumvents DAgger's need for ongoing expert availability through preference-based refinement.

Online RL with Offline datasets. The paradigm of initializing policies through offline pre-training followed by online fine-tuning has gained considerable traction, mirroring successes in supervised learning. Early contributions in this domain include model-based algorithms tailored for hybrid settings, such as the work by Ross & Bagnell (2012). After that, Xie et al. (2021) studied this hybrid RL setting and showing that offline data does not yield statistical improvements in tabular MDPs. This is different from our result, due to our expert's data. Recently, empirical RL algorithms designed to be effective in both offline and online contexts have been proposed, aiming to facilitate seamless offline-to-online fine-tuning (Rajeswaran et al., 2017; Hester et al., 2018; Nair et al., 2018; Vecerik et al., 2017; Lee et al., 2022; Ball et al., 2023). On the more theoretical side Song et al. (2023); Wagenmaker & Pacchiano (2023); Tang et al. (2023) proposes statistical approaches to efficiently combine offline and online datasets. Although these methods are related to our work, they assume access to numeric rewards during fine-tuning. Our approach eliminates this, assuming access to an expert trajectories dataset and preference-based online feedback.

Prior imitation-only approaches lack robustness outside the demonstration manifold; offline RL fine-tuning usually demands ground-truth rewards. Our work bridges these gaps by (i) proving that demonstration coverage plus a modest preference-query budget yields sharper high-probability regret bounds, and (ii) showing empirically that preference-guided exploration fixes blind spots with far fewer risky interactions than pure online RL.

3 Problem formulation

We address the challenge of learning optimal policies by combining information from two complementary sources: offline expert demonstrations and online preference feedback. In this hybrid learning paradigm, we first leverage a dataset $\mathbb{D}$ of trajectories collected from an expert policy to establish strong priors over the policy space. Then, we strategically utilize these priors to guide an online preference-based learning process, where an expert provides binary feedback comparing pairs of trajectories. This framework enables us to efficiently narrow the search space using offline demonstrations while refining our understanding of the expert's underlying preference model through targeted online queries. We aim to quantify how knowledge from offline demonstrations translates to improved regret bounds in the online preference learning phase.

Finite MDP setting (reward-free). Consider a finite-horizon Markov Decision Process (MDP) defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, H)$, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, $H \in \mathbb{N}$ is the horizon length, and $P = \{P_h\}_{h\in[H]}$ represents the time-dependent transition dynamics, with $P_h(\cdot|s,a)$ denoting the probability distribution over next states given state-action pair (s,a) at step h. A policy $\pi = \{\pi_h\}_{h\in[H]}$ consists of a collection of mappings $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ is the probability simplex over actions. A trajectory $\tau = \{(s_h, a_h)\}_{h\in[H]}$ is a sequence of state-action pairs generated by executing a policy π in the environment following dynamics P. We denote the space of all possible trajectories as $\mathcal{T}$. We assume the trajectories have a continuous distribution with respect to counting or Lebesgue measure. For ease of notation, we will write $\mathbb{P}_P^\pi$ for the density function of the trajectory distribution induced by policy π and dynamics P.

Offline demonstrations. We assume access to an offline dataset $\mathbb{D}_n^H = \{\tau_i\}_{i\in[n]}$ consisting of n independent trajectories of length H, where each $\tau_i \sim \mathbb{P}_{P^*}^{\pi^*}$. This represents an imitation learning framework where trajectories are generated by an expert policy π^* interacting with the true environment dynamics P^*.

Online preference queries. We formalize preference-based learning through feature embeddings and a probabilistic preference model (Christiano et al., 2017; Saha et al., 2023). For each trajectory $\tau \in \mathcal{T}$, we assume the existence of a trajectory embedding function $\phi : \mathcal{T} \to \mathbb{R}^d$ that is known to the learner. This creates a natural complementarity between our learning phases: while offline demonstrations provide raw trajectories that directly capture expert behavior, the embedding function transforms these complex sequences into a structured representation space that facilitates preference learning. The trajectory embedding function ϕ serves a critical purpose in our framework by enabling meaningful preference comparisons that would be difficult to perform on raw trajectories. This embedding approach provides a versatile framework that can accommodate various types of trajectory information. The flexibility of this representation allows our method to adapt to different domains and preference structures without changing the underlying learning algorithm. A policy π and dynamics P induce a distribution over trajectories, allowing us to define the policy embedding as the expected feature representation: $\phi^P(\pi) = \mathbb{E}_{\tau\sim\mathbb{P}_P^\pi}[\phi(\tau)]$.

In our work, we adopt two commonly used assumptions, bounded trajectory embeddings (Saha et al., 2023; Parker-Holder et al., 2020b) and bounded weight vectors (Filippi et al., 2010; Faury et al., 2020),

Assumption 3.1 (Bounded features). *For all $\tau \in \mathcal{T}$, the feature embeddings are bounded: $\|\phi(\tau)\|_2 \leq B$ for some known constant $B < \infty$.*

Assumption 3.2 (Bounded weights). *There exists an unknown weight vector $\mathbf{w}^* \in \{v \in \mathbb{R}^d : \|v\|_2 \leq W\}$ where the bound $W < \infty$ is known.*

Definition 1. *We measure the degree of non-linearity of the sigmoid σ over the parameter space (where σ' is the first derivative of σ) with*

$$\kappa := \sup_{\mathbf{x}\in B_B(d), \mathbf{w}\in B_S(d)} \frac{1}{\sigma'(\mathbf{w}^\top \mathbf{x})}.$$

We model the preference feedback through a Bradley-Terry model. Given two trajectories τ_1 and τ_2, the binary preference outcome $o_{1,2} \sim \text{Ber}(P)$ is modeled as:

$$\mathbb{P}(\tau_1 \succ \tau_2) = \mathbb{P}(o_{1,2} = 1 | \tau_1, \tau_2) = \sigma(\langle \phi(\tau_1) - \phi(\tau_2), \mathbf{w}^* \rangle),$$

where $\sigma(x) = (1 + e^{-x})^{-1}$ is the logistic function. This formulation corresponds to a latent utility model where the inner product $\langle \phi(\tau), \mathbf{w}^* \rangle$ represents the utility of trajectory τ.

From this model, we derive a score function for trajectories $s(\tau) = \langle \phi(\tau), \mathbf{w}^* \rangle$ and extend it to policies as $s^P(\pi) = \mathbb{E}_{\tau \sim \mathbb{P}^{\pi}_{P^*}}[s(\tau)]$, where P^* are the true transition dynamics. The preference between two policies π_1 and π_2 can now be written as follows: $\mathbb{P}(\pi_1 \succ \pi_2) = \sigma(\langle \phi^P(\pi_1) - \phi^P(\pi_2), \mathbf{w}^* \rangle)$. This represents an expected preference over the distribution of trajectories, and captures the average preference when comparing behaviors induced by different policies.

Offline estimation quality. For the offline phase, we measure the quality of estimation using distributional distance metrics in the space of trajectory distributions. Specifically, we will construct confidence sets in the form of Hellinger balls around our estimated density policy and dynamics. Notably, the Hellinger distance relates directly to the L^2 norm between square-root densities, enabling a geometric interpretation of our confidence sets as Euclidean balls in the space of density embeddings, with computational advantages over alternative divergences. The precise construction of these confidence sets and their properties will be detailed in the Section 4.

Online regret. We quantify our online learning phase's performance through regret measurement. In each round $t \in [T]$ of online learning, the agent selects policies π_t^1 and π_t^2, receives binary preference feedback $o_t \in \{0, 1\}$, and accumulates regret measured against the optimal policy. We specifically use the pseudo-regret with respect to the policy class Π as in Saha et al. (2023):

$$R_T^{\text{psr}} := \max_{\pi \in \Pi} \sum_{t=1}^{T} \frac{[2\phi^{P^*}(\pi) - \phi^{P^*}(\pi_t^1) - \phi^{P^*}(\pi_t^2)]^\top \mathbf{w}^*}{2} = \sum_{t=1}^{T} \frac{2s^{P^*}(\pi^*) - (s^{P^*}(\pi_t^1) + s^{P^*}(\pi_t^2))}{2},$$

where $\pi^* := \arg\max_{\pi \in \Pi} s(\pi)$.[1] All our performance guarantees will be expressed in terms of the MDP parameters (state space size $|\mathcal{S}|$, action space size $|\mathcal{A}|$, horizon length H), offline data quantity n, online interaction rounds T, and confidence level δ of the offline estimation – establishing a direct connection between offline data quality and online learning efficiency.

Notation. We denote $[H] = \{1, \ldots, H\}$ for $H \in \mathbb{N}$. For probability distributions P, Q, $H^2(P, Q)$ is the squared Hellinger distance and $\text{TV}(P, Q)$ the total variation distance. We denote $\mathbb{B}_2^d(R) := \{x \in \mathbb{R}^d : \|x\|_2 \leq R\}$ for the Euclidean ball of radius R, and for any $x \in \mathbb{R}^d$, we define $x^{\otimes 2} := xx^\top$ as the outer product.

4 Bridging offline behavioral cloning and online preference-based feedback

Our framework leverages offline expert demonstrations to improve the efficiency of online preference learning. The key insight is that we can use maximum likelihood estimation (MLE) on the offline dataset $\mathbb{D}_n^H$ to construct confidence sets in the policy space that likely contain the expert policy. This approach has two important features: First, by using the Hellinger distance, we obtain confidence sets that correspond to Euclidean norm balls in the space of square-root densities, providing a geometrically intuitive interpretation. The radius of this ball shrinks at a rate of $\mathcal{O}(1/\sqrt{n})$ with offline sample size, establishing a quantifiable relationship between offline data quantity and online learning efficiency. Second, we develop a technique to make this confidence set computable using only observed data, despite the theoretical formulation involving unknowns.

When we restrict the online phase of BRIDGE to sample only policies from this confidence set, we significantly reduce the number of expert preference queries needed compared to algorithms without

[1] Saha et al. (2023) showed that the standard preference-based regret formulation is equivalent up to constant factors, when $B, W \leq 1$.

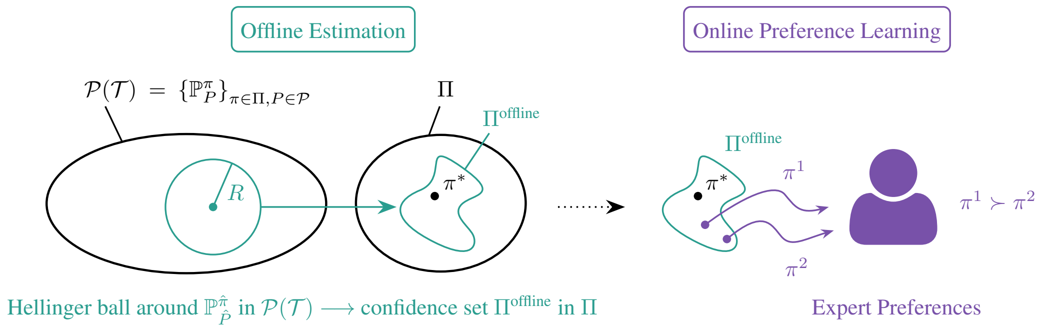

Offline Estimation
Online Preference Learning
$\mathcal{P}(\mathcal{T}) = \{\mathbb{P}^{\pi}_{P}\}_{\pi\in\Pi, P\in\mathcal{P}}$
Π
Π^{offline}
R
π^*
Π^{offline}
π^*
π^1
π^2
$\pi^1 \succ \pi^2$
Hellinger ball around $\mathbb{P}^{\hat{\pi}}_{\hat{P}}$ in $\mathcal{P}(\mathcal{T}) \longrightarrow$ confidence set Π^{offline} in Π
Expert Preferences

Figure 1: Overview of BRIDGE framework. Offline estimation derives estimators $\hat{\pi}$ and $\hat{P}$ using the dataset $\mathbb{D}_n^H$ and constructs a confidence set in trajectory distribution space $\mathcal{P}(\mathcal{T})$ as a Hellinger ball (left), which translates to the offline policy confidence set Π^{offline} in policy space Π likely to contain π^* (middle). The confidence set Π^{offline} is then handed to the online preference learning phase (right), where policies are sampled from within this set and presented to the expert for preference feedback.

offline data access. This hybrid approach effectively trades offline demonstrations for reduced online expert interaction. Geometrically, our confidence set has a clear interpretation as a Hellinger ball in the space of trajectory distributions. However, when mapped to the policy space, it forms a more complex shape due to the nonlinearity of the inverse functional mapping from distribution metrics to policy parameters. Figure 1 explains the relation between these quantities. In the rest of this section, we formally explain how BRIDGE uses offline behavioral cloning data to warm-start an online preference-based learning process.

4.1 Offline behavioral cloning and uncertainty estimators

We apply maximum likelihood estimation (MLE) on the offline dataset $\mathbb{D}_n^H$ of expert trajectories to obtain separate estimators $\hat{\pi}$, $\hat{P}$ for the optimal policy and transition model respectively. We then derive a confidence set around the estimated optimal policy $\hat{\pi}$, which constrains the policy search space in the second, online preference-based learning part of our method. Relevant corollaries and their proofs are presented in Appendix B. We start by making a standard realizability assumption.

Assumption 4.1 (Realizability)**.** *The optimal policy belongs to the policy class, $\pi^* \in \Pi$, and the true transition function belongs to the transition class, $P^* \in \mathcal{P}$.*

Policy estimation via log-loss Behavior Cloning. We define the *log-loss behavioral cloning estimator* as

$$\hat{\pi} = \arg\max_{\pi\in\Pi} \sum_{i\in[n]} \sum_{h\in[H]} \log(\pi_h(a_h^i|s_h^i)). \tag{1}$$

We characterize the estimation error in terms of the Hellinger distance between trajectory distributions in Corollary 19 in Appendix B.3, using concentration results by (Foster et al., 2024), cf. Appendix B.1.

Transition model estimation via Maximum Likelihood Estimation (MLE). Similarly, we define the *MLE transition estimator* as

$$\hat{P} = \arg\max_{P\in\mathcal{P}} \sum_{i\in[n]} \sum_{h\in[H]} \left(\log[P(s_{h+1}^i|s_h^i, a_h^i)] \right). \tag{2}$$

We describe the transition estimation quality in Corollary 23 in Appendix B.3.2, using maximum density likelihood concentration results by Foster et al. (2024).[2]

[2] Note that while we present results specifically for tabular, stochastic and stationary transitions, our framework readily adapts to other transition model classes by deriving appropriate covering number bounds using the general results in Appendix B.

Policy confidence set construction. We start by defining the *concentrability coefficient*, commonly encountered in offline RL literature (Chen & Jiang, 2019), as:

$$C(\hat{\pi}, \pi^*) = \sup_{t \in [H]} \sup_{(s,a) \in \mathcal{S} \times \mathcal{A} : d_{P^*}^{\pi^*,t}(s,a) > 0} \frac{d_{P^*}^{\hat{\pi},t}(s,a)}{d_{P^*}^{\pi^*,t}(s,a)}.$$

Intuitively, this coefficient measures how much the state-action visitation distribution of policy $\hat{\pi}$ can deviate from that of the expert policy π^* under the true dynamics P^*. To bound this quantity, we make the following standard assumption (Levine et al., 2020; Chen & Jiang, 2019) about minimum state-action visitation:

Assumption 4.2 (Minimum Visitation Probability)**.** *There exists a constant $\gamma_{min} > 0$ such that if a state-action-time tuple has a non-zero visitation probability under the optimal policy, that probability is bigger than this constant:*

$$\min_{(s,a,t): d_{P^*}^{\pi^*,t}(s,a) > 0} d_{P^*}^{\pi^*,t}(s,a) \geq \gamma_{min}. \tag{3}$$

Given this assumption, we can derive a deterministic bound on the concentrability coefficient:

Lemma 2 (Concentrability Coefficient Bound)**.** *Consider a policy estimator $\hat{\pi}$ satisfying*

$$H^2(\mathbb{P}_{P^*}^{\hat{\pi}}, \mathbb{P}_{P^*}^{\pi^*}) \leq R.$$

Then, under Assumption 4.2, the concentration coefficient is bounded by

$$C(\hat{\pi}, \pi^*) \leq 1 + \frac{2\sqrt{R}}{\gamma_{min}}.$$

This bound is key to our approach: it allows us to replace the unknown concentrability coefficient with a deterministic upper bound that depends only on our policy estimation error and the minimum visitation probability. We can now construct a practical confidence set as shown in the following lemma:

Lemma 3 (Offline Policy Confidence Set)**.** *Assume the following events hold:*

$$E_1 := \left\{ H^2(\mathbb{P}_{P^*}^{\hat{\pi}}, \mathbb{P}_{P^*}^{\pi^*}) \leq R_1(\delta_1) \right\}, \quad \textit{such that} \quad Prob(E_1) \geq 1 - \delta_1,$$
$$E_2 := \left\{ H^2(\mathbb{P}_{\hat{P}}^{\pi^*}, \mathbb{P}_{P^*}^{\pi^*}) \leq R_2(\delta_2) \right\}, \quad \textit{such that} \quad Prob(E_2) \geq 1 - \delta_2.$$

Then, under Assumption 4.2, the policy set

$$\Pi_{1-\delta}^{\textit{offline}} := \left\{ \pi : \sqrt{H^2(\mathbb{P}_{\hat{P}}^{\pi}, \mathbb{P}_{\hat{P}}^{\hat{\pi}})} \leq \sqrt{R_1} + \sqrt{R_2} \cdot \left(1 + \sqrt{\left(1 + \frac{2\sqrt{R_1}}{\gamma_{min}} \right) \cdot H} \right) \right\}$$

is a confidence set of level $1 - \delta = 1 - (\delta_1 + \delta_2)$, i.e.,

$$Prob(\pi^* \in \Pi_{1-\delta}^{\textit{offline}}) \geq 1 - (\delta_1 + \delta_2).$$

The key insight is that the concentrability coefficient now appears as a deterministic term in the confidence set radius, allowing us to construct a practical confidence region using only quantities that can be computed from offline data, along with our domain knowledge about the minimum visitation probability. This confidence set will be central to our online learning phase, providing a principled way to constrain the policy search space. By combining our tabular setting results from Corollaries 19 and 23 with the concentrability coefficient bound under Assumption 4.2, we can derive an explicit formula for the confidence set radius:

Theorem 4 (Offline Confidence Set Radius)**.** *Under the setting described above and Assumption 4.2, with $\delta_1 = \delta_2 = \delta/2$ and defining*

$$\alpha := \sqrt{4 \cdot |\mathcal{S}| \cdot \log(|\mathcal{A}| \cdot 2/\delta)},$$
$$\beta := \sqrt{4 \cdot |\mathcal{S}|^2 \cdot |\mathcal{A}| \cdot \log(nH \cdot 2/\delta)}.$$

The policy set

$$\Pi_{1-\delta}^{\textit{offline}} := \left\{ \pi : \sqrt{H^2(\mathbb{P}^{\pi}_{\hat{P}}, \mathbb{P}^{\hat{\pi}}_{\hat{P}})} \leq \textit{Radius} \right\},$$

is a confidence set of level $1-\delta$ containing π^* with probability at least $1-\delta$, where

$$\textit{Radius} = \frac{\alpha}{\sqrt{n}} + \frac{\beta}{\sqrt{n}} \cdot \left(1 + \sqrt{H \cdot \left(1 + \frac{2\alpha}{\gamma_{min} \cdot \sqrt{n}} \right)} \right).$$

This result provides the fundamental connection between offline data and online learning efficiency: the confidence set radius scales as $\mathcal{O}(1/\sqrt{n})$ with the offline sample size n. This inverse square root dependence means that as we collect more offline expert demonstrations, the confidence set shrinks, constraining the online policy search space more tightly. Since our online regret bounds will directly depend on the size of this confidence set, this establishes a quantifiable trade-off between offline data collection and online preference query efficiency, a key contribution of our work.

4.2 Online preference-based learning

In this section, we present how BRIDGE uses behavioral cloning estimation to guide online preference-based learning. Although we adapted the algorithm from Saha et al. (2023), our offline-to-online approach can be applied to other preference-based RL algorithms beyond BRIDGE.

Our online preference learning follows Saha et al. (2023), who adapted generalized linear models from parametric bandits (Filippi et al., 2010; Faury et al., 2020) to the preference-based RL setting. As in Saha et al. (2023) we compute a regularized maximum likelihood estimator $\mathbf{w}_t^{MLE}$ to learn the preference weight vector $\mathbf{w}^*$ from pairwise comparisons. However, since $\mathbf{w}_t^{MLE}$ may not satisfy assumption 3.2, as in Saha et al. (2023) we define the *data matrix* $\mathbf{V}_t$, which approximates the Fisher Information Matrix (the negative expected Hessian of the log-likelihood):

$$\mathbf{V}_t = \kappa\lambda\mathbf{I}_d + \sum_{\ell=1}^{t-1} (\phi(\tau_\ell^1) - \phi(\tau_\ell^2))^{\otimes 2}, g_t(\mathbf{w}) = \sum_{l=1}^{t-1} \sigma\left(\langle \phi(\tau_l^1) - \phi(\tau_l^2), \mathbf{w}\rangle\right)\left(\phi(\tau_l^1) - \phi(\tau_l^2)\right) + \lambda\mathbf{w}.$$

Then, the projected parameter, a constrained version of $\mathbf{w}_t^{MLE}$, is given by

$$\mathbf{w}_t^{proj} = \arg\min_{\mathbf{w} \in \mathcal{W}} \|g_t(\mathbf{w}) - g_t(\mathbf{w}_t^{MLE})\|_{\mathbf{V}_t^{-1}}. \tag{4}$$

This matrix $\mathbf{V}_t$ serves two purposes: First, defining a confidence ellipsoid

$$C_t(\delta) = \{\mathbf{w} : \|\mathbf{w} - \mathbf{w}_t^{proj}\|_{\mathbf{V}_t} \leq 2\kappa\beta_t(\delta)\}$$

containing $\mathbf{w}^*$ with high probability that is shaped by the likelihood curvature. Second, guiding exploration by quantifying uncertainty through $\|\cdot\|_{\mathbf{V}_t^{-1}}$, prioritizing directions with sparse information. This approach can be further strengthened by relating the empirical norm $\|\cdot\|_{\mathbf{V}_t}$ to an expected norm $\|\cdot\|_{\overline{\mathbf{V}}_t}$, where $\overline{\mathbf{V}}_t = \kappa\lambda\mathbf{I}_d + \sum_{\ell=1}^{t-1}(\phi^{\hat{P}_t}(\pi_\ell^1) - \phi^{\hat{P}_t}(\pi_\ell^2))^{\otimes 2}$. It can be shown that $\|\cdot\|_{\mathbf{V}_t}$ is approximately equivalent to $\|\cdot\|_{\overline{\mathbf{V}}_t}$ up to terms depending on the confidence bonus.

Our key contribution consists of connecting the offline estimation with the online learning process through two mechanisms: (i) leveraging the offline data to make a first estimation of the transition model, and (ii) constructing an initial dataset of policies that are plausible with the expert data. We refer the reader to (Appendix C) for a detailed explanation of the likelihood-based mechanisms inspired by online binary bandits that underpin this approach.

Transition Model Integration. We leverage our offline MLE transition estimate as the initialization for online learning. In the tabular setting, the MLE for transition probabilities from Eq. (2) is equivalent to a count-based estimator. As we collect online data, we update this estimator to combine both offline and online counts:

$$\hat{P}_t(s'|s,a) = \frac{N_{\text{off}}(s',s,a) + N_t(s',s,a)}{N_{\text{off}}(s,a) + N_t(s,a)},$$

where $N_{\text{off}}(s',s,a)$ counts transitions from state-action pair (s,a) to state s' in the offline dataset, $N_{\text{off}}(s,a)$ counts visits to (s,a), and $N_t(s',s,a)$ and $N_t(s,a)$ are the corresponding counts from the first t rounds of online interaction. We define our adapted bonus incorporating both offline and online data as:

$$\hat{B}_t(\pi,\eta,\delta) = \mathbb{E}_{\tau\sim\mathbb{P}^{\pi}_{\hat{P}_t}}\left[\sum_{h\in[H]} \min\left(2\eta, 4\eta\sqrt{\frac{U_h}{N_{\text{off}}(s_h,a_h) + N_t(s_h,a_h)}}\right)\right],$$

where $U_h = H\log(|\mathcal{S}||\mathcal{A}|) + \log\left(\frac{6\log(N_{\text{off}}(s_h,a_h)+N_t(s_h,a_h))}{\delta}\right)$. This adapts the bonus structure from Chatterji et al. (2021) to leverage our combined offline-online transition estimator.

Remark 5. *This integration approach, while effective, has potential for further improvement. First, it does not fully leverage the independence structure of the offline dataset, which could lead to tighter concentration bounds. Second, potential distribution shifts between offline and online phases are not explicitly modeled. These refinements represent promising directions for future research, though our primary focus remains on policy fine-tuning rather than optimal transition modeling.*

Algorithm 1 BRIDGE: Bounded Regret with Imitation Data and Guided Exploration

```
1: Input: offline dataset $\mathbb{D}_n^H$, no. of iterations $T$
2: Estimate transitions $\hat{P}$ via MLE (Eqn. (2)) and compute confidence set $\Pi^{\text{offline}}_{1-\delta}$ (Thm. 4)
3: Initialize $\hat{P}_1 \leftarrow \hat{P}$, $\overline{\mathbf{V}}_1 \leftarrow \lambda\mathbf{I}_d$        ▷ Initialize model and data matrix
4: for $t = 1,\dots,T$ do
5:     Compute $\mathbf{w}^{\text{proj}}_t$ via constrained MLE (Eqn. (4))
6:     Define policy set $\Pi_t$ based on $\Pi^{\text{offline}}_{1-\delta}$ and $\mathbf{w}^{\text{proj}}_t$ (Lemma 7)
7:     $(\pi^1_t,\pi^2_t) \leftarrow \arg\max_{\pi^1,\pi^2\in\Pi_t}\{\gamma_t\cdot\|\phi^{\hat{P}_t}(\pi^1) - \phi^{\hat{P}_t}(\pi^2)\|_{\overline{\mathbf{V}}_t^{-1}}$
           $+ \hat{B}_t(\pi^1, 2WB, \delta^{online}) + \hat{B}_t(\pi^2, 2WB, \delta^{online})\}$
8:     Sample trajectories $\tau^1_t \sim \mathbb{P}^{\pi^1_t}_{\hat{P}_t}$, $\tau^2_t \sim \mathbb{P}^{\pi^2_t}_{\hat{P}_t}$ and obtain preference $o_t = \mathbb{I}(\tau^1_t \succ \tau^2_t)$
9:     Update matrix $\overline{\mathbf{V}}_{t+1} \leftarrow \overline{\mathbf{V}}_t + (\phi^{\hat{P}_t}(\pi^1_t) - \phi^{\hat{P}_t}(\pi^2_t))^{\otimes 2}$ and model $\hat{P}_{t+1}$
10: end for
11: return Best policy from $\Pi_T$ using final weight estimate $\mathbf{w}^{\text{proj}}_t$
```

Feature Moment Bounds. We constrain policy selection to our offline-derived confidence set $\Pi^{\text{offline}}_{1-\delta}(\Pi)$. This enables us to bound the difference between expected feature representations of policies, which directly impacts the uncertainty quantification in our online learning phase.

Lemma 6 (Feature Moment Bounds). *Let X be a random variable on measurable space $(\mathcal{X},\tilde{\mathcal{X}})$ and $f : \mathcal{X} \to \mathbb{R}^d$ be a bounded function such that $\|f(x)\|_2 \le B < \infty$ for all $x \in \mathcal{X}$. For probability distributions P, Q that are absolutely continuous with respect to the Lebesgue measure, if $H^2(P,Q) \le R$, then*

$$\|\mathbb{E}_{X\sim P}[f(X)] - \mathbb{E}_{X\sim Q}[f(X)]\|_2 \le 2\sqrt{2}\cdot B\cdot\sqrt{R}.$$

This lemma provides a crucial connection between the Hellinger distance of trajectory distributions and the distance between their expected feature representations (embeddings) by setting $f = \phi$. In our preference-based learning framework, we apply this result to bound the elements of the expected feature covariance matrix $\overline{\mathbf{V}}_t$. The trace $\mathrm{tr}(\overline{\mathbf{V}}_t) = \sum_{t'<t} \|\phi(\pi_1^{t'}) - \phi(\pi_2^{t'})\|_2^2$ represents total variance, which Lemma 6 controls via our confidence set construction.

At the start of our online learning process ($t = 0$), for policies π_0^1 and π_0^2 that belong to our offline-derived confidence set $\Pi_{1-\delta}^{\text{offline}}$ from Theorem 4, the Hellinger distance between their induced trajectory distributions is bounded by the confidence set radius. Applying Lemma 6 with $f = \phi$, our trajectory embedding function, we obtain:

$$\|\phi^{\hat{P}_0}(\pi_0^1) - \phi^{\hat{P}_0}(\pi_0^2)\|_2 \leq 4\sqrt{2} \cdot B \cdot \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

This result has profound implications for our regret analysis. By constraining policies to our confidence set, we effectively control the variance of feature differences, allowing us to replace the naive bound $\|\phi^{\hat{P}_t}(\pi_1^t) - \phi^{\hat{P}_t}(\pi_2^t)\|_2 \leq 2B$ with our tighter bound that decreases with the offline sample size n. As online learning progresses, the transition model $\hat{P}_t$ improves through additional data collection. Importantly, this improvement in the transition model only strengthens our bound. The exact form of this improvement depends on the concentration properties of the online estimator and is formalized in our final regret proof.

Based on these bounds, we can now define a policy confidence set that contains the optimal policy π^* with high probability while accounting for both estimation uncertainty and exploration bonuses:

Lemma 7 (Online Policy Confidence Set)**.** *Let Π_t be the set of policies defined as*

$$\begin{aligned}\Pi_t := \Big\{ \pi \in \Pi_{1-\delta}^{\textit{offline}} \;\Big|\; \forall \pi' \in \Pi_{1-\delta}^{\textit{offline}} :&\\ &\left\langle \phi^{\hat{P}_t}(\pi) - \phi^{\hat{P}_t}(\pi'), \mathbf{w}_t^{\textit{proj}} \right\rangle + \gamma_t \cdot \|\phi^{\hat{P}_t}(\pi) - \phi^{\hat{P}_t}(\pi')\|_{\overline{\mathbf{V}}_t^{-1}} \\ &+ \hat{B}_t(\pi, 2WB, \delta') + \hat{B}_t(\pi', 2WB, \delta') \geq 0 \Big\},\end{aligned}$$

where $\delta' = \frac{\delta^{\textit{online}}}{2|\mathcal{A}|^{|\mathcal{S}|}}$. With probability at least $1 - \delta^{\textit{online}}$, the optimal policy π^* remains in Π_t for all $t \in [T]$, where $\delta^{\textit{online}}$ has been scaled appropriately via union bound to account for the separate probabilistic events in the offline confidence set, transition model estimation, and parameter estimation components. The definition of the confidence radius multiplier γ_t is provided in **??**.

The complete algorithm is described in Algorithm 1.

4.3 Theoretical guarantees

We can now state the following regret bound for BRIDGE. The proof is shown in Appendix D.

Theorem 8 (Regret Bound with Offline-Enhanced Exploration)**.** *Let n be the number of offline demonstrations with minimum visitation probability $\gamma_{\min} > 0$ for state-action pairs. With probability at least $1 - \delta$, the regret of* BRIDGE *is bounded by*

$$R_T \leq 2 \cdot \underbrace{\gamma_T}_{\textit{Term 1}} \cdot \underbrace{\sqrt{T \cdot \log\left(1 + \frac{\tilde{\mathcal{O}}\left(B^2 \cdot H \cdot |S|^2 \cdot \min\left\{\frac{T}{n}, \log(T)\right\} + \frac{T \cdot |S| \cdot B^2 \cdot \sqrt{|A| \cdot H}}{\sqrt{n \cdot \gamma_{\min}}}\right)}{d}\right)}}_{\textit{Term 2}}$$

$$+ \underbrace{\tilde{\mathcal{O}}\left(H|\mathcal{S}|\sqrt{\frac{|\mathcal{A}|TH}{n \cdot \gamma_{\min}}} + \frac{H^{5/2}WB\sqrt{T}}{\sqrt{n \cdot \gamma_{\min}}} + H^2WB \cdot \sqrt{T} \cdot \frac{|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/4}}{n^{1/4}}\right)}_{\textit{Term 3}},$$

where

$$\gamma_T = \tilde{\mathcal{O}}\left((\kappa + BW)\sqrt{d\log(T)} + H^2WB|\mathcal{S}| \cdot \sqrt{\min\left\{\log(T), \frac{T}{n \cdot \gamma_{\min}}\right\}} + \sqrt{H^2WB} \right),$$

we have set $\epsilon = \frac{1}{T}$ to optimize the bound and we omit logarithmic factors (excluding T's) in $\tilde{\mathcal{O}}$.

From this regret bound we can observe that as $n \to \infty$ with fixed $\gamma_{\min} > 0$: (i) Term 1 approaches $\tilde{\mathcal{O}}((\kappa + BW)\sqrt{d\log(T)} + \sqrt{HWB})$; (ii) Term 2, the logarithm, approaches $\log(1) = 0$; (iii) in Term 3, all components approach zero. The overall regret bound exhibits a $\sqrt{T}$ dependence as in Saha et al. (2023). However, this results in a regret bound that can be made arbitrarily small with sufficiently high-quality offline data, changing the complexity of regret analysis without having access to an offline expert dataset. This result helps in closing the gap between empirical results in applying RL in real-world scenarios and theoretical works.

4.4 Numerical simulations

We provide initial simulation results based on a practical implementation of our algorithm. As baselines we used Foster et al. (2024)'s behavioral cloning (BC) and Saha et al. (2023)'s PbRL algorithms (for which no implementations are publicly available). We used the `StarMDP` and `Gridworld` environments described in Pace et al. (2024). Appendix F gives more details on our experiments. Figure 2 shows that with as little as 2 offline, optimal trajectories, BRIDGE prunes the policy set and converges to the optimal policy faster than PbRL.

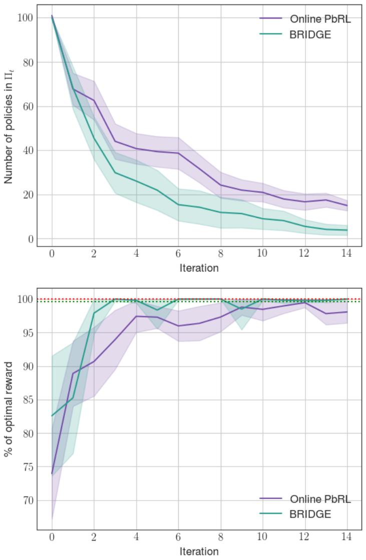


Figure 2: Performance comparison of our new algorithm with offline BC (Foster et al., 2024) and online PbRL Saha et al. (2023). The dotted red and green lines are the expected return of the optimal and BC policies respectively.

5 Conclusions

We present BRIDGE[3], a novel algorithm for fine-tuning BC policies using online PbRL. Our approach is motivated by the practical challenges of deploying RL in real-world settings, where reward specification is difficult and exploration is risky. By combining these two feedbacks, we construct confidence sets that constrain the policy space and guide safe, sample-efficient learning. We provide the first theoretical regret bound for this hybrid learning paradigm, showing that offline data reduces regret through a shrinking uncertainty radius. Our analysis builds on recent advances in BC and PbRL, and crucially integrates them into a unified regret framework with provable benefits. Our work opens new directions for interactive learning systems that can safely and efficiently improve with human input, even in the absence of explicit reward signals.

[3] Code available on `https://github.com/pfriedric/bridge`.

References

Philip J Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. In *International Conference on Machine Learning*, pp. 1577–1594. PMLR, 2023.

Daniel Brown, Russell Coleman, Ravi Srinivasan, and Scott Niekum. Safe imitation learning via fast bayesian reward inference from preferences. In *International Conference on Machine Learning*, pp. 1165–1177. PMLR, 2020.

Niladri Chatterji, Aldo Pacchiano, Peter Bartlett, and Michael Jordan. On the theory of reinforcement learning with once-per-episode feedback. *Advances in Neural Information Processing Systems*, 34: 3401–3412, 2021.

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

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Proc. of Advances in Neural Information Processing Systems*, volume 2017-Decem, pp. 4300–4308, 2017.

Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*, 2019.

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pp. 3052–3060. PMLR, 2020.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. *Advances in neural information processing systems*, 23, 2010.

Dylan J Foster, Adam Block, and Dipendra Misra. Is behavior cloning all you need? understanding horizon in imitation learning. *arXiv preprint arXiv:2407.15007*, 2024.

Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, et al. Deep q-learning from demonstrations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Ilya Kostrikov, Ofir Nachum, Vikas Tomar, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Seunghyun Lee, Younggyo Seo, Kimin Lee, Pieter Abbeel, and Jinwoo Shin. Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble. In *Conference on Robot Learning*, pp. 1702–1712. PMLR, 2022.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Ashvin Nair, Bob McGrew, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Overcoming exploration in reinforcement learning with demonstrations. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 6292–6299. IEEE, 2018.

Ashvin Nair, Gal Dalal, Abhishek Gupta, and Sergey Levine. Awac: Accelerating online reinforcement learning with offline datasets. In *Conference on robot learning*, pp. 102–121. PMLR, 2020.

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.

Aldo Pacchiano, Jack Parker-Holder, Yunhao Tang, Krzysztof Choromanski, Anna Choromanska, and Michael Jordan. Learning to score behaviors for guided policy optimization. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7445–7454. PMLR, 2020. URL https://proceedings.mlr.press/v119/pacchiano20a.html.

Alizée Pace, Bernhard Schölkopf, Gunnar Rätsch, and Giorgia Ramponi. Preference elicitation for offline reinforcement learning. *arXiv preprint arXiv:2406.18450*, 2024.

Seohong Park, Kevin Frans, Sergey Levine, and Aviral Kumar. Is value learning really the main bottleneck in offline rl? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Jack Parker-Holder, Aldo Pacchiano, Krzysztof Choromanski, and Stephen Roberts. Effective diversity in population based reinforcement learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020a. Curran Associates Inc. ISBN 9781713829546.

Jack Parker-Holder, Aldo Pacchiano, Krzysztof M Choromanski, and Stephen J Roberts. Effective diversity in population based reinforcement learning. *Advances in Neural Information Processing Systems*, 33:18050–18062, 2020b.

Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1, 1988.

Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *arXiv preprint arXiv:1709.10087*, 2017.

Stéphane Ross and J Andrew Bagnell. Agnostic system identification for model-based reinforcement learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1905–1912, 2012.

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

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

Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

Yuda Song, Yifei Zhou, Ayush Sekhari, Drew Bagnell, Akshay Krishnamurthy, and Wen Sun. Hybrid RL: Using both offline and online data can make RL efficient. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=yyBis80iUuU.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Chen Tang, Ben Abbatematteo, Jiaheng Hu, Rohan Chandra, Roberto Martín-Martín, and Peter Stone. Deep reinforcement learning for robotics: A survey of real-world successes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28694–28698, 2025.

Dengwang Tang, Rahul Jain, Botao Hao, and Zheng Wen. Efficient online learning with offline datasets for infinite horizon mdps: A bayesian approach. *arXiv preprint arXiv:2310.11531*, 2023.

Andrea Tirinzoni, Ahmed Touati, Jesse Farebrother, Mateusz Guzek, Anssi Kanervisto, Yingchen Xu, Alessandro Lazaric, and Matteo Pirotta. Zero-shot whole-body humanoid control via behavioral foundation models. *arXiv preprint arXiv:2504.11054*, 2025.

Sara A. van de Geer. Applications of empirical process theory. 2000. URL https://api.semanticscholar.org/CorpusID:123051755.

Mel Vecerik, Todd Hester, Jonathan Scholz, Fumin Wang, Olivier Pietquin, Bilal Piot, Nicolas Heess, Thomas Rothörl, Thomas Lampe, and Martin Riedmiller. Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. *arXiv preprint arXiv:1707.08817*, 2017.

Andrew Wagenmaker and Aldo Pacchiano. Leveraging offline data in online reinforcement learning. In *International Conference on Machine Learning*, pp. 35300–35338. PMLR, 2023.

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

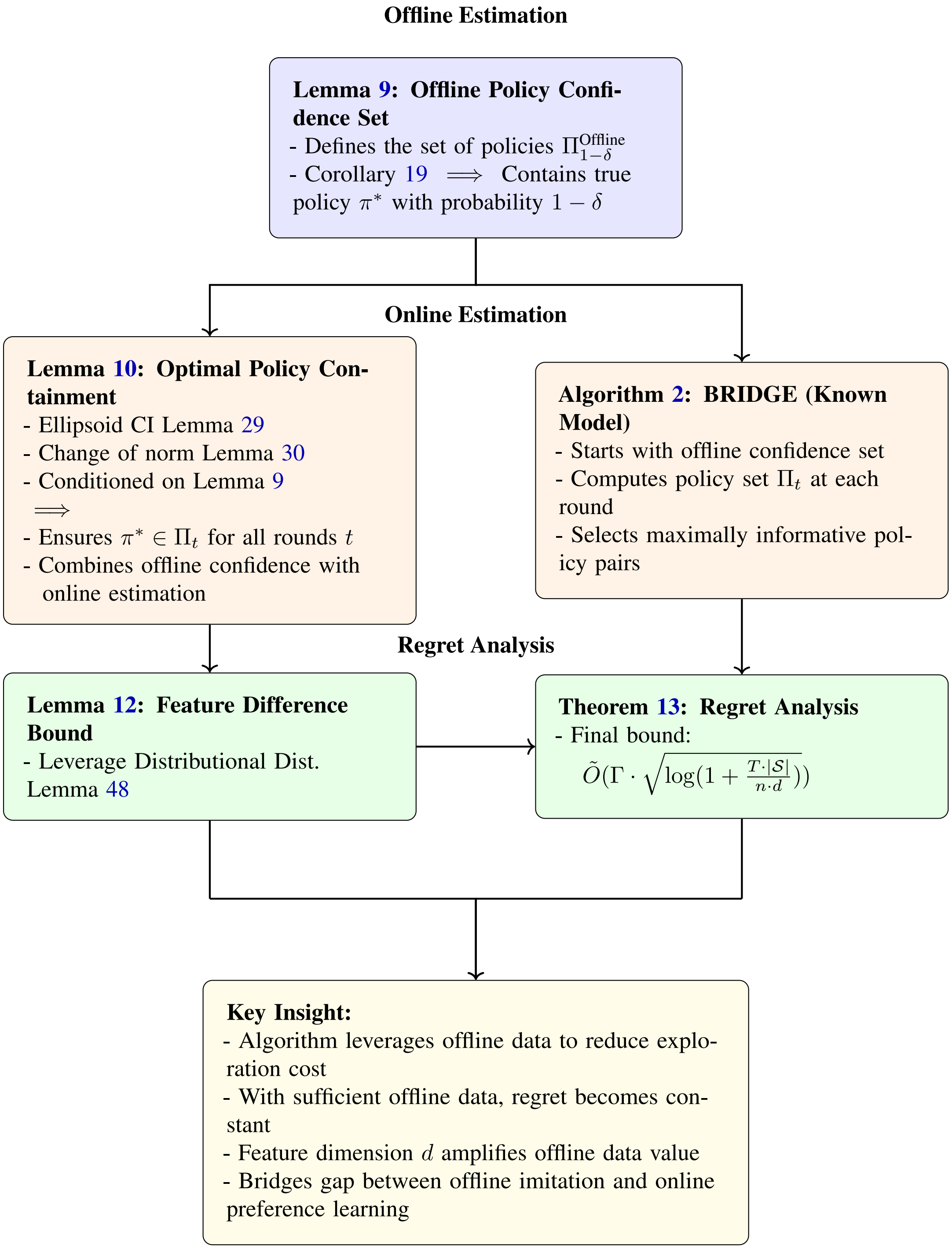
Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021.

Tong Zhang. From ε-entropy to kl-entropy: Analysis of minimum information complexity density estimation. *Annals of Statistics*, 34:2180–2210, 2006. URL https://api.semanticscholar.org/CorpusID:14152942.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Simplified Setup for Understanding Regret Analysis



Lemma 9: Offline Policy Confidence Set
- Defines the set of policies $\Pi_{1-\delta}^{\text{Offline}}$
- Corollary 19 $\implies$ Contains true policy π^* with probability $1-\delta$

Lemma 10: Optimal Policy Containment
- Ellipsoid CI Lemma 29
- Change of norm Lemma 30
- Conditioned on Lemma 9

$\implies$
- Ensures $\pi^* \in \Pi_t$ for all rounds t
- Combines offline confidence with online estimation

Algorithm 2: BRIDGE (Known Model)
- Starts with offline confidence set
- Computes policy set Π_t at each round
- Selects maximally informative policy pairs

Lemma 12: Feature Difference Bound
- Leverage Distributional Dist. Lemma 48

Theorem 13: Regret Analysis
- Final bound: $\tilde{O}(\Gamma \cdot \sqrt{\log(1 + \frac{T \cdot |\mathcal{S}|}{n \cdot d}}))$

Key Insight:
- Algorithm leverages offline data to reduce exploration cost
- With sufficient offline data, regret becomes constant
- Feature dimension d amplifies offline data value
- Bridges gap between offline imitation and online preference learning

In this section, we propose an analysis of the regret under a simplified setting: The underlying dynamic P^* is known. This idea is to help the reader understand how the construction of the confidence set over the policies from the offline learning estimation helps to reduce the number of policies to draw from in the online learning setting, without being overwhelmed by the estimation of the transitions. In this setting, it is then clear what part of our methods applies to the policies. The goal is to prepare the reader for the proof of our algorithm BRIDGE in Appendix D.

A.1 Setup for Known Dynamics

A.1.1 Offline Estimation with Known Dynamics

Assume we get the offline data $\mathbb{D}_n^H = \{\tau_i\}_{i\in[n]}$. The underlying object describing the trajectories is a Finite MDP Reward Free setting as in the main paper. Assume that the set of possible policies is stationary and deterministic. Then under the fact that the underlying dynamic is known, the confidence set from Theorem 4 reduces to the following, by direct application of Corollary 19, i.e., setting the radius around the MLE estimate π_{MLE} from Equation (7).

We formalize this into the following lemma:

Lemma 9 (Offline Policy Confidence Set under Known Dynamics)**.** *Let $\hat{\pi}$ be the log-loss BC estimator defined in Equation (7).*
The policy set

$$\Pi_{1-\delta}^{\textit{Offline}} := \left\{\pi : H(\mathbb{P}_{P^*}^{\pi}, \mathbb{P}_{P^*}^{\hat{\pi}}) \leq \sqrt{\frac{6\cdot|\mathcal{S}|\cdot\log(|\mathcal{A}|\cdot\delta^{-1})}{n}}\right\}$$

contains π^* with probability at least $1-\delta$.

Proof. Note that by symmetry

$$H(\mathbb{P}_{P^*}^{\pi^*}, \mathbb{P}_{P^*}^{\hat{\pi}}) = H(\mathbb{P}_{P^*}^{\hat{\pi}}, \mathbb{P}_{P^*}^{\pi^*})$$

Then the result follows from Corollary 19. □

A.1.2 Online Learning with Known Dynamics

Here we adapt our algorithm BRIDGE to the setting with known dynamics. This means we adapt the approach from Saha et al. (2023) under known dynamics to constrain the set of policies to choose from to our confidence set described in the previous section.

First, since the transitions are known, we define for this section:

$$\phi^{P^*}(\pi) := \phi(\pi) = \mathbb{E}_{\tau\sim\mathbb{P}_{P^*}^{\pi}}[\phi(\tau)]$$

We also define the expected data matrix $\overline{V}_t^{P^*}$ under the true transition dynamics P^* as follows (see Appendix C for an overview of results about data matrices):

$$\overline{V}_t^{P^*} = \kappa\lambda\mathbf{I}_d + \sum_{\ell=1}^{t-1}\left(\phi(\pi_\ell^1) - \phi(\pi_\ell^2)\right)\left(\phi(\pi_\ell^1) - \phi(\pi_\ell^2)\right)^\top$$

Then we define the set of policies to draw from as:

$$\Pi_t := \Big\{\, \pi \in \Pi_{1-\delta}^{\text{Offline}} \;\Big|\, \forall \pi' \in \Pi_{1-\delta}^{\text{Offline}} : \\ \left\langle \phi(\pi) - \phi(\pi'), \mathbf{w}_t^{\text{proj}} \right\rangle + \gamma_t \cdot \|\phi(\pi) - \phi(\pi')\|_{(\overline{\mathbf{V}}_t^{P^*})^{-1}} \geq 0 \,\Big\}$$

where $\gamma_t := 2\kappa\beta_t(\delta) + \alpha_{d,T}(\delta)$ and $\alpha_{d,T}(\delta)$ is defined as in Lemma 30.

Lemma 10 (Optimal Policy Containment)**.** *Conditioned on $E_{w^*} \cap E_{\overline{V}_T^{P^*}} \cap E_{\textit{offline}}$ where:*

- *E_{w^*} is the event defined in Lemma 29*
- *$E_{\overline{V}_T^{P^*}}$ is the event defined in Lemma 30*
- *$E_{\textit{offline}} := \{\pi^* \in \Pi_{1-\delta}^{\textit{Offline}}\}$*

then

$$\pi^* \in \Pi_t \quad \forall t \in [T]$$

Proof. This follows directly from Lemma 2 in Saha et al. (2023). We adapt the probability parameter δ to account for the additional condition that $\pi^* \in \Pi_{1-\delta}^{\text{Offline}}$, which holds with probability at least $1-\delta$ according to Lemma 27. □

We now present the adapted version of BRIDGE for the known transition model:

Algorithm 2 BRIDGE: Bounded Regret with Imitation Data and Guided Exploration (**Known Model**)

```
Input: Offline dataset $\mathbb{D}_n^H$, time horizon $T$, true dynamics $P^*$
Compute confidence set $\Pi_{1-\delta}^{\text{Offline}}$ using Lemma 9
Initialize $\overline{\mathbf{V}}_1^{P^*} \leftarrow \kappa\lambda\mathbf{I}_d$          ▷ Initialize data matrix
for $t = 1, \dots, T$ do
    Compute $\mathbf{w}_t^{\text{proj}}$ via constrained MLE (Equation (4))
    Define policy set $\Pi_t$ based on $\Pi_{1-\delta}^{\text{Offline}}$ and $\mathbf{w}_t^{\text{proj}}$
    $(\pi_t^1, \pi_t^2) \leftarrow \arg\max_{\pi^1,\pi^2 \in \Pi_t} \{\|\phi(\pi^1) - \phi(\pi^2)\|_{(\overline{\mathbf{V}}_t^{P^*})^{-1}}\}$
    Sample trajectories $\tau_t^1 \sim \mathbb{P}_{P^*}^{\pi_t^1}$, $\tau_t^2 \sim \mathbb{P}_{P^*}^{\pi_t^2}$ and obtain preference $o_t = \mathbb{I}(\tau_t^1 \succ \tau_t^2)$
    Update matrix: $\overline{\mathbf{V}}_{t+1}^{P^*} \leftarrow \overline{\mathbf{V}}_t^{P^*} + (\phi(\pi_t^1) - \phi(\pi_t^2))(\phi(\pi_t^1) - \phi(\pi_t^2))^\top$
end for
return Best policy from $\Pi_T$ using final weight estimate $\mathbf{w}_T^{\text{proj}}$
```

A.2 Regret Analysis: BRIDGE (Known Model)

We now present a regret analysis of the BRIDGE algorithm under known transition. We start by stating the following lemma:

Lemma 11. *The regret of* BRIDGE *under known dynamic is upper bounded as follow:*

$$R_T \leq 4 \cdot (\beta_T(\delta) + \alpha_{T,d}(\delta)) \cdot \sum_{t \in [T]} \|\phi(\pi_t^1) - \phi(\pi_t^2)\|_{(\overline{V}_t^{P^*})^{-1}}$$

Proof. For ease of notation we define $\delta^{*,1}\phi := \phi(\pi^*) - \phi(\pi^1)$. First we bound the instantenous regret

$$\begin{aligned}
2r_t &= \langle \delta^{*,1}\phi, w^* \rangle + \langle \delta^{*,2}\phi, w^* \rangle \\
&\leq \langle \delta^{*,1}\phi, \mathbf{w}_t^{\text{proj}} \rangle + \langle \delta^{*,2}\phi, \mathbf{w}_t^{\text{proj}} \rangle + \|w^* - \mathbf{w}_t^{\text{proj}}\|_{\overline{V}_t^{P^*}} \left(\|\delta^{*,1}\phi\|_{(\overline{V}_t^{P^*})^{-1}} + \|\delta^{*,2}\phi\|_{(\overline{V}_t^{P^*})^{-1}} \right) \\
&\underbrace{\leq}_{Lemma~30} \langle \delta^{*,1}\phi, \mathbf{w}_t^{\text{proj}} \rangle + \langle \delta^{*,2}\phi, \mathbf{w}_t^{\text{proj}} \rangle + (2\kappa\beta_t(\delta) + \alpha_{T,d}(\delta)) \cdot \left(\|\delta^{*,1}\phi\|_{(\overline{V}_t^{P^*})^{-1}} + \|\delta^{*,2}\phi\|_{(\overline{V}_t^{P^*})^{-1}} \right)
\end{aligned}$$

Then notice that choosing π_t^1, π_t^2 as $\arg\max$ together with the fact that $\pi^* \in \Pi_t$ Lemma 10 yield

$$2r_t \leq \langle \delta^{*,1}\phi, \mathbf{w}_t^{\text{proj}} \rangle + \langle \delta^{*,2}\phi, \mathbf{w}_t^{\text{proj}} \rangle + (2\kappa\beta_t(\delta) + \alpha_{T,d}(\delta)) \cdot \left(\|\delta^{1,2}\phi\|_{(\overline{V}_t^{P^*})^{-1}} \right)$$

Next using the fact that $\pi_t^1, \pi_t^2, \pi^* \in \Pi_t$ we have the following constraints

$$\begin{aligned} &\langle \delta^{i,*}\phi, \mathbf{w}_t^{\text{proj}} \rangle + \gamma_t \|\delta^{*,i}\phi\|_{(\overline{V}_t^{P^*})^{-1}} \geq 0 \quad i \in \{1,2\} \\ \Leftrightarrow &\langle \delta^{*,i}\phi, \mathbf{w}_t^{\text{proj}} \rangle \leq \gamma_t \|\delta^{*,i}\phi\|_{(\overline{V}_t^{P^*})^{-1}} \quad i \in \{1,2\} \end{aligned}$$

yielding

$$2r_t \leq (2\kappa\beta_t(\delta) + \alpha_{T,d}(\delta))\|\delta^{1,2}\phi\|_{(\overline{V}_t^{P^*})^{-1}} \leq 4(2\kappa\beta_T(\delta) + \alpha_{T,d}(\delta)) \cdot \|\delta^{1,2}\phi\|_{(\overline{V}_t^{P^*})^{-1}}$$

Hence

$$R_T = \sum_{t\in[T]} r_t \leq 2 \cdot \beta_T(\delta) + \alpha_{T,d}(\delta)) \cdot \sum_{t\in[T]} \|\delta^{1,2}\phi\|_{(\overline{V}_t^{P^*})^{-1}}$$

□

The remaining step in our analysis is to bound the term:

$$\sum_{t\in[T]} \|\phi(\pi_t^1) - \phi(\pi_t^2)\|_{(\overline{V}_t^{P^*})^{-1}} = \sum_{t\in[T]} \left\| \mathbb{E}_{\tau\sim\mathbb{P}_{P^*}^{\pi_t^1}}[\phi(\tau)] - \mathbb{E}_{\tau\sim\mathbb{P}_{P^*}^{\pi_t^2}}[\phi(\tau)] \right\|_{(\overline{V}_t^{P^*})^{-1}}$$

A simple approach would be to use **??**, which states that the feature map ϕ is bounded in ℓ_2-norm by B. However, our offline confidence set construction in Lemma 9 provides a more powerful result: policies in our set have distributions that are close not only in Hellinger distance but also in the resulting feature expectations.

This is precisely why we formulated our confidence set constraint using the square root of the squared Hellinger distance - it yields a bound on the L_2 norm of distribution differences. Through Lemma 48, we can translate bounds on Hellinger distance into bounds on the difference of feature expectations in the ℓ_2-norm.

We formalize this connection in the following lemma:

Lemma 12 (Feature Difference Bound Under Offline Constraints - Corrected)**.** *For policies $\pi_t^1, \pi_t^2 \in \Pi_{1-\delta}^{\textit{Offline}}$ selected by our algorithm at each round $t \in [T]$, the sum of feature differences measured in the data matrix norm is bounded as:*

$$\sum_{t\in[T]} \|\phi(\pi_t^1) - \phi(\pi_t^2)\|_{(\overline{V}_t^{P^*})^{-1}} \leq \sqrt{2d \cdot \log\left(1 + \frac{192B^2T|\mathcal{S}|\log(|\mathcal{A}| \cdot \delta^{-1})}{n \cdot d \cdot \lambda}\right)}$$

where d is the feature dimension, B is the feature norm bound, $|\mathcal{S}|$ and $|\mathcal{A}|$ are the state and action space sizes, and n is the number of offline samples.

Proof. First note

$$\sum_{t\in[T]} \|\phi(\pi_t^1) - \phi(\pi_t^2)\|_{(\overline{V}_t^{P^*})^{-1}} \leq \sqrt{T \cdot \sum_{t\in[T]} \|\phi(\pi_t^1) - \phi(\pi_t^2)\|^2_{(\overline{V}_t^{P^*})^{-1}}}$$

then notice the inequality

$$u \leq 2\log(1+u) \quad u \geq 1 \implies \sum_{t\in[T]} \|\delta^{1,2}\phi\|^2_{(\overline{V}_t^{P^*})^{-1}} \leq 2 \cdot \sum_{t\in[T]} \log(1 + \|\delta^{1,2}\phi\|^2_{(\overline{V}_t^{P^*})^{-1}})$$

Using the definition of $\overline{V}_t^{P^*}$, we have

$$\begin{aligned}
\overline{V}_{t+1}^{P^*} &= \lambda \cdot I_{d\times d} + \sum_{i\in[t]} (\phi(\pi_i^1) - \phi(\pi_i^2))(\phi(\pi_i^1) - \phi(\pi_i^2))^T \\
&= \overline{V}_t^{P^*} + (\phi(\pi_t^1) - \phi(\pi_t^2))(\phi(\pi_t^1) - \phi(\pi_t^2))^T \\
&= (\overline{V}_t^{P^*})^{1/2}\left(I + (\overline{V}_t^{P^*})^{-1/2}(\phi(\pi_t^1) - \phi(\pi_t^2))(\phi(\pi_t^1) - \phi(\pi_t^2))^T(\overline{V}_t^{P^*})^{-1/2}\right)(\overline{V}_t^{P^*})^{1/2}
\end{aligned}$$

Using properties of determinant:

$$\begin{aligned}
\det(\overline{V}_{t+1}^{P^*}) &= \det(\overline{V}_t^{P^*}) \cdot \det(I + (\overline{V}_t^{P^*})^{-1/2}(\phi(\pi_t^1) - \phi(\pi_t^2))(\phi(\pi_t^1) - \phi(\pi_t^2))^T(\overline{V}_t^{P^*})^{-1/2}) \\
&= \det(\overline{V}_t^{P^*}) \cdot (1 + \|\phi(\pi_t^1) - \phi(\pi_t^2)\|^2_{(\overline{V}_t^{P^*})^{-1}}) \\
&= \det(V_0) \cdot \prod_{s\in[t]} (1 + \|\phi(\pi_s^1) - \phi(\pi_s^2)\|^2_{(\overline{V}_s^{P^*})^{-1}}) \\
&\Leftrightarrow \\
&\log\left[\frac{\det(\overline{V}_{t+1}^{P^*})}{\det(V_0)}\right] = \sum_{s\in[t]} \log(1 + \|\phi(\pi_s^1) - \phi(\pi_s^2)\|^2_{(\overline{V}_s^{P^*})^{-1}})
\end{aligned}$$

We have for the determinant:

$$\det(\overline{V}_{t+1}^{P^*}) = \prod_{i\in[d]} \lambda_i \le \left(\frac{1}{d} \cdot \text{Tr}\{\overline{V}_{t+1}^{P^*}\}\right)^d$$

Using linearity of trace:

$$\begin{aligned}
\text{Tr}\{\overline{V}_{t+1}^{P^*}\} &= \text{Tr}\{\lambda I\} + \sum_{s\in[t]} \text{Tr}\{(\phi(\pi_s^1) - \phi(\pi_s^2))(\phi(\pi_s^1) - \phi(\pi_s^2))^T\} \\
&= d \cdot \lambda + \sum_{s\in[t]} \|\phi(\pi_s^1) - \phi(\pi_s^2)\|_2^2
\end{aligned}$$

Applying the corrected bound from Lemma 48:

$$\begin{aligned}
\|\phi(\pi_t^1) - \phi(\pi_t^2)\|_2^2 &\le (2\sqrt{2} \cdot B \cdot \sqrt{H^2(\mathbb{P}_{P^*}^{\pi_t^1}, \mathbb{P}_{P^*}^{\pi_t^2})})^2 \\
&\le 8B^2 \cdot \frac{24 \cdot |\mathcal{S}| \cdot \log(|\mathcal{A}| \cdot \delta^{-1})}{n} \\
&= \frac{192B^2|\mathcal{S}|\log(|\mathcal{A}| \cdot \delta^{-1})}{n}
\end{aligned}$$

Using this tighter bound in our trace calculation:

$$\begin{aligned}
\text{Tr}\{\overline{V}_{t+1}^{P^*}\} &\le d \cdot \lambda + t \cdot \frac{192B^2|\mathcal{S}|\log(|\mathcal{A}| \cdot \delta^{-1})}{n} \\
&= d\lambda\left(1 + \frac{192B^2 t|\mathcal{S}|\log(|\mathcal{A}| \cdot \delta^{-1})}{n \cdot d \cdot \lambda}\right)
\end{aligned}$$

Hence:

$$\begin{aligned}
\log\left[\frac{\det(\overline{V}_{t+1}^{P^*})}{\det(V_0)}\right] &\leq d\cdot\log\left(\frac{\mathrm{Tr}\{\overline{V}_{t+1}^{P^*}\}}{d}\right)\\
&= d\cdot\log\left(\lambda\left(1+\frac{192B^2t|\mathcal{S}|\log(|\mathcal{A}|\cdot\delta^{-1})}{n\cdot d\cdot\lambda}\right)\right)\\
&= d\cdot\log(\lambda)+d\cdot\log\left(1+\frac{192B^2t|\mathcal{S}|\log(|\mathcal{A}|\cdot\delta^{-1})}{n\cdot d\cdot\lambda}\right)
\end{aligned}$$

Since $\det(V_0)=\lambda^d$, the first logarithmic term cancels out:

$$\log\left[\frac{\det(\overline{V}_{t+1}^{P^*})}{\det(V_0)}\right] = d\cdot\log\left(1+\frac{192B^2t|\mathcal{S}|\log(|\mathcal{A}|\cdot\delta^{-1})}{n\cdot d\cdot\lambda}\right)$$

Therefore:

$$\begin{aligned}
\sum_{t\in[T]}\|\phi(\pi_t^1)-\phi(\pi_t^2)\|^2_{(\overline{V}_t^{P^*})^{-1}} &\leq 2\cdot\log\left[\frac{\det(\overline{V}_{T+1}^{P^*})}{\det(V_0)}\right]\\
&\leq 2d\cdot\log\left(1+\frac{192B^2T|\mathcal{S}|\log(|\mathcal{A}|\cdot\delta^{-1})}{n\cdot d\cdot\lambda}\right)
\end{aligned}$$

Taking the square root:

$$\sum_{t\in[T]}\|\phi(\pi_t^1)-\phi(\pi_t^2)\|_{(\overline{V}_t^{P^*})^{-1}} \leq \sqrt{2d\cdot\log\left(1+\frac{192B^2T|\mathcal{S}|\log(|\mathcal{A}|\cdot\delta^{-1})}{n\cdot d\cdot\lambda}\right)}$$

□

Theorem 13 (Regret Analysis for BRIDGE under Known Model)**.** *Let $\delta\leq 1/e$ and $\lambda\geq\frac{B}{\kappa}$. Then, with probability at least $1-\delta$, the expected regret of Algorithm 2 is bounded by:*

$$R_T\leq(2\kappa\beta_T(\delta)+\alpha_{d,T}(\delta))\sqrt{2d\cdot\log\left(1+\frac{192B^2T|\mathcal{S}|\log(|\mathcal{A}|\cdot\delta^{-1})}{n\cdot d\cdot\lambda}\right)}$$

In asymptotic notation, this becomes:

$$R_T=O\left(\left(W\sqrt{\kappa B}+WB\right)d\log(TB/\kappa\delta)\sqrt{\log\left(1+\frac{T|\mathcal{S}|}{n\cdot d}\right)}\right)$$

where the probability parameter δ accounts for the events

$$\begin{array}{lcl}
E_{w^*} & \rightarrow & \textit{Lemma } 29\\
E_{\overline{V}_T^{P^*}} & \rightarrow & \textit{Lemma } 30\\
E_{\textit{offline}} := \{\pi^*\in\Pi_{1-\delta}^{\textit{Offline}}\} & \rightarrow & \textit{Lemma } 9
\end{array}$$

Remark 14. *This result demonstrates a significant improvement over Saha et al. (2023)'s bound of $O\left(\left(W\sqrt{\kappa B}+WB\right)d\log(TB/\kappa\delta)\sqrt{T}\right)$. The key advantage lies in the term $\sqrt{\log(1+\frac{T|\mathcal{S}|}{n})}$, which approaches zero as $n\to\infty$, potentially yielding constant regret.*

A.3 Practical Regret Analysis with Fixed Offline Data

For a fixed offline dataset of size n, our regret bound scales with horizon T as:

$$R_T = O\left(\Gamma \cdot \sqrt{\log\left(1 + \frac{CT|\mathcal{S}|}{n \cdot d}\right)}\right)$$

where $\Gamma = (W\sqrt{\kappa B} + WB)d\log(TB/\kappa\delta)$. This bound reveals three distinct regimes:

1. **Small T Regime** ($T|\mathcal{S}| \ll n \cdot d$): Using $\log(1+x) \approx x$ for small x:

$$R_T = O\left(\Gamma \cdot \sqrt{\frac{T|\mathcal{S}|}{n \cdot d}}\right) = O\left(\Gamma \cdot \sqrt{T} \cdot \frac{|\mathcal{S}|^{1/2}}{\sqrt{n \cdot d}}\right)$$

2. **Transition Regime** ($T|\mathcal{S}| \approx n \cdot d$):

$$R_T = O(\Gamma) = O\left((W\sqrt{\kappa B} + WB)d\log(TB/\kappa\delta)\right)$$

3. **Large T Regime** ($T|\mathcal{S}| \gg n \cdot d$):

$$R_T = O\left(\Gamma \cdot \sqrt{\log(T)}\right)$$

These regimes highlight two key insights: (1) with sufficient offline data ($n = \Omega(\frac{T|\mathcal{S}|}{d})$), regret dramatically improves from $O(\sqrt{\log(T)})$ to $O(1)$ in the dependence on T; and (2) feature dimension d amplifies the value of offline data, allowing the same regret reduction with $\sqrt{d}$ times less data. This explains why high-dimensional problems may benefit more significantly from offline data.

As n increases, regret transitions from logarithmic ($O(\log(T))$) to sublinear ($O(\sqrt{T/n})$) and eventually approaches $O(1)$ when $n \gg \frac{T|\mathcal{S}|}{d}$. In the limiting case where $n \to \infty$, exploration becomes unnecessary, and regret is bounded only by statistical error in the offline estimation.

B Offline estimation

B.1 Maximum Likelihood for Density Estimation

In this section, we present Maximum Likelihood Estimation (MLE) for density estimation that forms the foundation of our concentration results. While these results are presented more extensively in Foster et al. (2024), we include them here for completeness and readability.

The analysis of MLE relies on standard concentration techniques following the well-established work of van de Geer (2000) and Zhang (2006), enhanced by new Freedman-type concentration inequalities developed in Foster et al. (2024) (Appendix B).

The key proof strategy connects MLE analysis to information-theoretic measures via Rényi divergence of order $1/2$. Specifically, the approach bounds expressions of the form $-n \cdot \log(\mathbb{E}_{z\sim g^*}[e^{\frac{1}{2}\log(g(z)/g^*(z))}])$, which equals $\frac{n}{2} \cdot D_{1/2}(g\|g^*)$. This term is bounded using Freedman-type inequalities for adapted sequences, which provide high-probability bounds of the form $\sum_{t=1}^{T'} -\log(\mathbb{E}_{t-1}[e^{-X_t}]) \leq \sum_{t=1}^{T'} X_t + \log(\delta^{-1})$. When combined with union bounds over ε-nets, this yields tight concentration results for the entire function class. The approach also leverages connections to Hellinger distance through the identity $H^2(g, g^*) = 1 - \int \sqrt{g(z)g^*(z)}dz$, providing geometrically interpretable guarantees.

To handle infinite classes, we introduce a tailored notion of covering number for log-loss:

Definition 15 (Log-Covering Number). *For a class $\mathcal{G} \subset \Delta(\mathcal{X})$, the class $\mathcal{G}' \subset \mathcal{X}$ is an $\epsilon-$cover if for all $g \in \mathcal{G}$, there exists $g' \in \mathcal{G}'$ such that $\forall x \in \mathcal{X}$*

$$\log(g(x)/g'(x)) \leq \epsilon$$

The size of such cover is defined by $\mathcal{N}_{\log}(\mathcal{G}, \epsilon)$.

Consider the data $\mathbb{D}_n = \{x_i\}_{i\in[n]}$ consisting of i.i.d copies of $x \sim g^*$ where $g^* \in \Delta(\mathcal{X})$. We have a class $\mathcal{G} \subseteq \Delta(\mathcal{X})$ that may or may not contain g^*. The density MLE estimator is defined as

$$\hat{g} = \arg\max_{g\in\mathcal{G}} \sum_{i\in[n]} \log(g(x_i)) \tag{5}$$

Lemma 16 (Maximum Likelihood Estimator Bound). *The maximum likelihood estimator in Eq. Equation (5) has that with probability at least $1-\delta$,*

$$H^2(\hat{g}, g^*) \leq \inf_{\varepsilon>0}\left\{\frac{6\log(2\mathcal{N}_{\log}(\mathcal{G},\varepsilon)/\delta^{-1})}{n} + 4\varepsilon\right\} + 2\inf_{g\in\mathcal{G}}\log(1 + D_{\chi^2}(g^*\|g))$$

In particular, if $\mathcal{G}$ is finite, the maximum likelihood estimator satisfies

$$H^2(\hat{g}, g^*) \leq \frac{6\log(2|\mathcal{G}|/\delta^{-1})}{n} + 2\inf_{g\in\mathcal{G}}\log(1 + D_{\chi^2}(g^*\|g))$$

Note that the term $\inf_{g\in\mathcal{G}}\log(1 + D_{\chi^2}(g^*\|g))$ corresponds to misspecification error, and is zero if $g^* \in \mathcal{G}$.

B.2 MLE Objective of Dataset of Independent Trajectories

Given a data set of reward free trajectories $\mathbb{D}_n^H = \{\tau_i\}_{i\in[H]}$ of n trajectories of length H where $\{\tau_i\} \sim_{i.i.d} \tau \sim \mathbb{P}_{P^*}^{\pi^*}$. The distribution $\mathbb{P}_{P^*}^{\pi^*}$ is assumed to be continuous w.r.t to Lebesque measure. It is characterized by the policy density $\pi = \{\pi_i\}_{i\in[H]} \in \Pi$ and the stationary transition density $P = \mathcal{P}$ where $\Pi, \mathcal{P}$ characterize the policy and transition density spaces. The log-likelihood of the set with for a policy π and a transition P reads:

$$l_n(\pi, P) = \frac{1}{n}\sum_{i\in[n]} \log\left[P(s_1^i)\cdot \pi_1(a_1^i, s_1^i)\prod_{1<j\leq H} P(s_j^i|s_{j-1}^i, a_{j-1}^i)\pi_j(a_j^i|s_j^i)\right]$$

The maximum likelihood objective over the density class $\{\mathbb{P}_P^\pi\}_{\pi\in\Pi, P\in\mathcal{P}}$ for the dataset $\mathbb{D}_n^H$

$$\arg\max_{\pi\in\Pi, P\in\mathcal{P}} \sum_{i\in[n]}\sum_{j\in[H]}\Big(\log[\pi_i(a_j^i|s_j^i])\Big) + \sum_{i\in[n]}\sum_{j=0}^{H}\Big(\log[P(s_{j+1}^i|s_j^i, a_j^i)]\Big) \tag{6}$$

B.3 Concentration Bounds

In this section we provide concentration bounds for the MLE estimators of the policies and the transition model, as well as for our notion of concentrability coefficient. The important takeaway is that the control of the error, i.e., the decay of these concentration bounds depends only on values known to the user, which will allow us to compute confidence policy sets based on these bounds.

B.3.1 Policy estimation

Define the log-loss behavioral cloning estimator for dataset $\mathbb{D}_n^H$ as described in B.2 as

$$\hat{\pi} = \arg\max_{\pi\in\Pi} \sum_{i\in[n]}\sum_{h\in[H]} \log(\pi_h(a_h^i|s_h^i)) \tag{7}$$

which is from Equation (6) equivalent to performing maximum density estimation over the density class $\{\mathbb{P}_{P^*}^\pi\}_{\pi\in\Pi}$. Similar to definition 15 (Foster et al., 2024) define the following

Definition 17 (Policy Covering Number)**.** *For a class* $\Pi \subset \{\pi_h : \mathcal{S} \to \Delta(\mathcal{A})\}$*, we say that* $\Pi' \subset \{\pi_h : \mathcal{S} \to \Delta(\mathcal{A})\}$ *is an* $\epsilon-$*cover if for all* $\pi \in \Pi$ *there exists* $\pi' \in \Pi'$ *such that*

$$\log\left(\frac{\pi_h(a|s)}{\pi'_h(a|s)}\right) \leq \epsilon \quad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$$

We denote the size of the smallest such cover as $\mathcal{N}_{pol}(\Pi, \epsilon)$

We state the following theorem from (Foster et al., 2024, Appendix C):

Theorem 18 (Generalization bound for logloss-BC)**.** *The Logloss BC estimator Eq:7 satisfies with probability* $\geq 1 - \delta$

$$H^2(\mathbb{P}^{\hat{\pi}}_{P^*}, \mathbb{P}^{\pi^*}_{P^*}) \leq \inf_{\epsilon} \left\{ \frac{6 \log(2\mathcal{N}_{pol}(\Pi, \epsilon/H)\delta^{-1})}{n} + \epsilon \right\}$$

in particular, if Π *is finite*

$$H^2(\mathbb{P}^{\hat{\pi}}_{P^*}, \mathbb{P}^{\pi^*}_{P^*}) \leq \frac{6 \cdot \log(2 \cdot |\Pi| \cdot \delta^{-1})}{n}$$

Proof. See (Foster et al., 2024, Appendix C). □

Corollary 19 (Deterministic Stationary Tabular Policies)**.** *If* $\Pi = \Pi^D_S$ *i.e the set of deterministic tabular policies the log-loss BC estimator Eq 7 has that with probability at least* $1 - \delta$

$$H^2(\mathbb{P}^{\hat{\pi}}_{P^*}, \mathbb{P}^{\pi^*}_{P^*}) \leq \frac{6 \cdot |\mathcal{S}| \cdot \log(|\mathcal{A}| \cdot \delta^{-1})}{n}$$

Proof. We have $|\Pi^D_S| = |\mathcal{A}|^{|\mathcal{S}|}$ □

In the case we don't have deterministic but stochastic policies, we need to determine $\log(\mathcal{N}_{pol}(\Pi_S, \epsilon))$. This can be accomplished using a discretization argument, where we create a finite ϵ-net that approximates the continuous space of stochastic policies within the desired error tolerance.

B.3.2 Transition model estimation

Here we can give a similar argument as for the policy log loss BC estimator. We define the following estimator

$$\hat{P} = \arg\max_{P \in \mathcal{P}} \sum_{i \in [n]} \sum_{j=0}^{H} \left(\log[P(s^i_{j+1}|s^i_j, a^i_j)] \right) \tag{8}$$

which is from Eq. 6 equivalent to performing maximum density estimation over the density class $\{\mathbb{P}^{\pi^*}_P\}_{P \in \mathcal{P}}$. Similarly, we define the following notion of covering

Definition 20 (Stationary Transition Log Covering Number)**.** *For a class of stationary transition probability functions* $\mathcal{P} \subset \{P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})\}$ *we define that* $\mathcal{P}' \subset \{P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})\}$ *is an* ϵ*-cover if for all* $P \in \mathcal{P}$ *there exists* $P' \in \mathcal{P}'$ *such that*

$$\log\left(\frac{P(s'|s,a)}{P'(s'|s,a)}\right) \leq \epsilon \quad \forall (s', s) \in \mathcal{S}, a \in \mathcal{A}$$

We denote the smallest such cover by $\mathcal{N}_{trans}(\mathcal{P}, \epsilon)$

Assumption B.1 (Realisability of Transition)**.** *We assume the true transition density to be in the model class i.e* $P^* \in \mathcal{P}$

We can now give a similar guarantee as for the log loss policy estimate but for the transition estimate

Theorem 21 (Generalisation Bound for MLE Transition Estimator)**.** *The MLE transition estimator of Eq 8 satisfies with probability at least* $1-\delta$

$$H^2(\mathbb{P}^{\pi^*}_{\hat{P}}, \mathbb{P}^{\pi^*}_{P^*}) \leq \inf_{\epsilon} \left\{ \frac{6\log(2\mathcal{N}_{trans}(\Pi, \epsilon/H)\delta^{-1})}{n} + \epsilon \right\}$$

Proof. Given a valid ϵ-cover of $\mathcal{P}$ from definition 20 we have

$$\log\left(\frac{\mathbb{P}^{\pi}_{P}}{\mathbb{P}^{\pi}_{P'}}\right) = \sum_{h=1}^{H} \log\left(\frac{P(s_{h+1}|s_h, a_h)}{P'(s_{h+1}|s_h, a_h)}\right) \leq \epsilon \cdot H$$

this means that we get a valid $\epsilon \cdot H$ cover for the trajectory density class. The bound follow be a direct application of Lemma 16. □

Lemma 22 (Transition Covering Number: Stationary Tabular Stochastic Transitions)**.** *For a class of stationary transition probability functions* $\mathcal{P} \subset \{P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ *where* $|\mathcal{S} \times \mathcal{A}|$ *is finite (tabular MDP), the* ϵ*-cover from Definition 20 satatisfies:*

$$\log(\mathcal{N}_{trans}(\mathcal{P}, \epsilon)) \leq |\mathcal{S}| \cdot |\mathcal{A}| \cdot (|\mathcal{S}| - 1) \cdot \log\left(\frac{1}{\epsilon} + 1\right)$$

Proof. The proof follow a standard geometric discretization argument for finite class of function (see Chapter 5 Wainwright (2019)) . For a given $\epsilon > 0$ we construct a geometric grid:

$$\mathcal{G}_\epsilon = \left\{\delta, \delta \cdot \exp(\epsilon/2), \delta\exp(\epsilon), \delta\exp(3\epsilon/2), \ldots, \delta\exp(k\epsilon/2)\right\}$$

where $\delta > 0$ is the minimum probability and k is chosen such that the grid represents a discretization of the continuous interval $[0, 1]$ i.e

$$\delta\exp(k\epsilon/2) \implies k \geq \frac{2\log(1/\delta)}{\epsilon}$$

Thus the grid size is at most:

$$|\mathcal{G}_\epsilon| \leq \left\lceil \frac{2\log(1/\delta)}{\epsilon} \right\rceil + 1$$

For each state action pairs (a, s) define $P(s_i|s, a) = p_i$ for $i \in |\mathcal{S}|$. Note that for the first $1, \ldots, |\mathcal{S}|-1$ there exist $q_i := P'(s_i|s, a) \in \mathcal{G}_\epsilon$ that satisfies by construction

$$\exp(-\epsilon/2) \leq \frac{p_i}{q_i} \leq \exp(\epsilon/2)$$

For the last state $i = |\mathcal{S}|$, we need to determine $q_{|\mathcal{S}|}$ close enough to $p_{|\mathcal{S}|}$.
Let define $S_q := \sum_i^{|\mathcal{S}-1|} q_i$ and $S_p := \sum_i^{|\mathcal{S}-1|} p_i$ we have the constraint

$$p_{|\mathcal{S}|} = 1 - S_p$$
$$q_{|\mathcal{S}|} = 1 - S - q$$

From the bound on the first $1 - |\mathcal{S}|$ elements we have

$$S_q \exp(-\epsilon/2) \leq S_p \leq S_q \exp(\epsilon/2)$$

For the ratio of the last probability

$$\frac{p_{|\mathcal{S}|}}{q_{|\mathcal{S}|}} = \frac{1 - S_p}{1 - S_q}$$

we have the following condition such that we have $\frac{p_{|\mathcal{S}|}}{q_{|\mathcal{S}|}} \geq \exp(-\epsilon)$

$$S_q \leq \frac{1 - \exp(-\epsilon)}{\exp(\epsilon/2) - \exp(-\epsilon)}$$

Similarly for the upper bound we have the condition

$$S_q \leq \frac{\exp(\epsilon) - 1}{\exp(\epsilon) - \exp(-\epsilon/2)}$$

Combining both constraints we have

$$S_q \leq \min\left\{\frac{\exp(\epsilon) - 1}{\exp(\epsilon) - \exp(-\epsilon/2)}, \frac{\exp(\epsilon) - 1}{\exp(\epsilon) - \exp(-\epsilon/2)}\right\}$$

By Taylor approximation this boils down to

$$S_q \leq \frac{2}{3}$$

Hence we select only the combination of points that satisfies

$$\frac{2}{3} \leq S_q \leq 1 - \delta$$

It remains to count the number of point we have in our cover i.e the first $|\mathcal{S} - 1|$ that satisfies our constrains

$$(\text{number of grid points})^{|\mathcal{S}-1|} \leq |\mathcal{G}_\epsilon|^{|\mathcal{S}|-1} \leq \left(\left\lceil \frac{2\log(1/\delta)}{\epsilon} \right\rceil + 1\right)^{|\mathcal{S}|-1}$$

Across all state action pair and taking the logarithm

$$\log(\mathcal{N}_{trans}(\mathcal{P}, \epsilon)) \leq |\mathcal{S}||\mathcal{A}|(\mathcal{S} - 1)\log\left(\left\lceil \frac{2\log(1/\delta)}{\epsilon} \right\rceil + 1\right)$$

choosing $\delta = O(\epsilon)$ yield the result. □

Corollary 23 (Stochastic, stationary, tabular transition setting)**.** *For finite $|\mathcal{S} \times \mathcal{A}|$ (tabular setting) and assuming the transition density class to be stochastic and stationary we have with probability at least $1 - \delta$*

$$H^2(\mathbb{P}^{\pi^*}_{\hat{P}}, \mathbb{P}^{\pi^*}_{P^*}) \leq \mathcal{O}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|\log(nH\delta^{-1})}{n}\right)$$

where for the theoretical optimal constant $C_{theory} \approx 6$

Proof. From our lemma on the covering number of transition functions, we have:

$$\log \mathcal{N}_{\text{trans}}(\mathcal{P}, \varepsilon/H) \leq |\mathcal{S}||\mathcal{A}|(|\mathcal{S}| - 1)\log\left(\frac{H}{\varepsilon} + 1\right)$$

For large $\frac{H}{\varepsilon}$, we can approximate:

$$\log\left(\frac{H}{\varepsilon} + 1\right) \approx \log\left(\frac{H}{\varepsilon}\right)$$

Substituting this into our bound:

$$
\begin{aligned}
H^2(\mathbb{P}^{\pi^*}_{\hat{P}}, \mathbb{P}^{\pi^*}_{P^*}) &\leq \inf_{\varepsilon > 0} \left\{ \frac{6\log(2) + 6|\mathcal{S}||\mathcal{A}|(|\mathcal{S}|-1)\log\left(\frac{H}{\varepsilon}\right) + 6\log(\delta^{-1})}{n} + \varepsilon \right\} \\
&= \inf_{\varepsilon > 0} \left\{ \frac{6\log(2) + 6D\log(H) - 6D\log(\varepsilon) + 6\log(\delta^{-1})}{n} + \varepsilon \right\}
\end{aligned}
$$

where $D = |\mathcal{S}||\mathcal{A}|(|\mathcal{S}|-1)$ for brevity.

To find the optimal ε, we differentiate with respect to ε and set to zero:

$$
\begin{aligned}
\frac{d}{d\varepsilon}\left[\frac{6\log(2) + 6D\log(H) - 6D\log(\varepsilon) + 6\log(\delta^{-1})}{n} + \varepsilon\right] &= -\frac{6D}{n\varepsilon} + 1 = 0 \\
&\Rightarrow \varepsilon_{\text{opt}} = \frac{6D}{n} = \frac{6|\mathcal{S}||\mathcal{A}|(|\mathcal{S}|-1)}{n}
\end{aligned}
$$

Substituting this optimal value back:

$$
\begin{aligned}
H^2(\mathbb{P}^{\pi^*}_{\hat{P}}, \mathbb{P}^{\pi^*}_{P^*}) &\leq \frac{6\log(2) + 6D\log(H) - 6D\log\left(\frac{6D}{n}\right) + 6\log(\delta^{-1})}{n} + \frac{6D}{n} \\
&= \frac{6\log(2) + 6D\log(H) - 6D\log(6D) + 6D\log(n) + 6\log(\delta^{-1}) + 6D}{n} \\
&= \frac{6\log(2) + 6\log(\delta^{-1}) + 6D\log\left(\frac{nH}{6D}\right) + 6D}{n}
\end{aligned}
$$

For large state spaces where $|\mathcal{S}| - 1 \approx |\mathcal{S}|$, and defining $\tilde{D} = |\mathcal{S}|^2|\mathcal{A}|$, this becomes:

$$
H^2(\mathbb{P}^{\pi^*}_{\hat{P}}, \mathbb{P}^{\pi^*}_{P^*}) \leq \frac{6\log(2) + 6\log(\delta^{-1}) + 6\tilde{D}\log\left(\frac{nH}{6\tilde{D}}\right) + 6\tilde{D}}{n}
$$

For large n and $\tilde{D}$, the dominant term is $\frac{6\tilde{D}\log(nH)}{n}$, and we can combine the logarithmic terms to get:

$$
H^2(\mathbb{P}^{\pi^*}_{\hat{P}}, \mathbb{P}^{\pi^*}_{P^*}) = O\left(\frac{|\mathcal{S}|^2|\mathcal{A}|\log(nH\delta^{-1})}{n}\right)
$$

Note that the constant 6 appears in the full derivation. This completes the proof. □

B.3.3 Concentrability Coefficient Upper Bound

Definition 24 (Concentrability Coefficient)**.** *We define the following quantity as the "concentrability coefficient":*

$$
C(\hat{\pi}, \pi^*) = \sup_{t\in[H]} \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}: d^{\pi^*,t}_{P^*}(s,a)>0} \frac{d^{\pi,t}_{P^*}(s,a)}{d^{\pi^*,t}_{P^*}(s,a)}
$$

which measures the maximum ratio between the state-action distributions induced by policies $\hat{\pi}$ and π^* under the true dynamics P^*.

Assumption B.2 (Minimum Visitation Probability)**.** *There exists a constant $\gamma_{min} > 0$ such that for all state-action-time tuples with non-zero probability under the optimal policy:*

$$
\min_{(s,a,t): d^{\pi^*,t}_{P^*}(s,a)>0} d^{\pi^*,t}_{P^*}(s,a) \geq \gamma_{min}
$$

Lemma 25 (Concentrability Coefficient Bound)**.** *Consider a policy estimator $\hat{\pi}$ satisfying*

$$H^2(\mathbb{P}^{\hat{\pi}}_{P^*}, \mathbb{P}^{\pi^*}_{P^*}) \leq R$$

Then, under Assumption 4.2, the concentration coefficient is bounded by:

$$C(\hat{\pi}, \pi^*) \leq 1 + \frac{2\sqrt{R}}{\gamma_{min}}$$

Proof. We will proceed by upper bounding the numerator using the condition on the Hellinger distance followed by lower bounding with concentration the denominator.
For the upper bound note that

$$\sup_{a,s} |d^{\hat{\pi},t}_{P^*} - d^{\pi^*,t}_{P^*}| = 2 \cdot TV(d^{\hat{\pi},t}_{P^*}, d^{\pi^*,t}_{P^*})$$

Recalling that the state-action distribution $d^{\pi,t}_{P^*}(s,a)$ is the marginal distribution of the trajectory distribution at time step t. Explictly:

$$d^{\pi,t}_{P^*}(s,a) = \int_{\tau_{-t}} \mathbb{P}^{\pi}_{P^*}(\tau)\, d\tau_{-t} =: \mathbb{P}^{\pi}_{P^*}(s_t = s, a_t = a)$$

where τ_{-t} denotes all time steps in the trajectory except for time t, and $\mathbb{P}^{\pi}_{P^*}(\tau)$ is the probability of trajectory τ under policy π and dynamics P^*.
Hence

$$\begin{aligned} TV(d^{\hat{\pi},t}_{P^*}, d^{\pi^*,t}_{P^*}) &= 2 \cdot TV(\mathbb{P}^{\hat{\pi}}_{P^*}(s_t = s, a_t = a), \mathbb{P}^{\pi^*}_{P^*}(s_t = s, a_t = a)) \\ &\leq 2 \cdot TV(\mathbb{P}^{\hat{\pi}}_{P^*}, \mathbb{P}^{\pi^*}_{P^*}) \\ &\leq 2 \cdot \sqrt{H^2(\mathbb{P}^{\hat{\pi}}_{P^*}, \mathbb{P}^{\pi^*}_{P^*})} \leq 2\sqrt{R} \end{aligned}$$

together with

By Assumption 1, we have a lower bound on the minimum state-action visitation probability:

$$\min_{(s,a,t): d^{\pi^*,t}_{P^*}(s,a) > 0} d^{\pi^*,t}_{P^*}(s,a) \geq \gamma_{min}$$

Finally, we combine the upper bound on the numerator and the lower bound from Assumption 1 to get $\forall (a,s,t) \quad \text{s.t} \quad d^{\pi^*,t}_{P^*}(a,s) > 0$:

$$\begin{aligned} C(\hat{\pi}, \pi^*) &= 1 + \frac{\sup_{a,s,t} |d^{\hat{\pi},t}_{P^*}(s,a) - d^{\pi^*,t}_{P^*}(s,a)|}{\inf_{a,s,t} d^{\pi^*,t}_{P^*}(s,a)} \\ &\leq 1 + \frac{2\sqrt{R}}{\inf_{a,s,t} d^{\pi^*,t}_{P^*}(s,a)} \\ &\leq 1 + \frac{2\sqrt{R}}{\gamma_{min}} \end{aligned}$$

This completes the proof, giving us a deterministic bound on the concentration coefficient that depends on the Hellinger distance between trajectory distributions and the minimum visitation probability of the optimal policy. □

B.3.4 Confidence Set Construction

In this section we will derive a distributional confidence set on the trajectory space in the form of a hellinger ball, accounting for the error of the MLE density estimates $\hat{\pi}$ and $\hat{P}$. We start by presenting the following in between result

Lemma 26 (Technical Results). *Assume finite state and action pair: $\mathcal{S} \times \mathcal{A}$. The following upper bound is true $\forall \pi \in \Pi$ with π^* being the true policy and $P^*, \hat{P}$ being the true and estimated transition models:*

$$H^2(\mathbb{P}^{\pi}_{\hat{P}}, \mathbb{P}^{\pi}_{P^*}) \leq H \cdot C(\pi, \pi^*) \cdot H^2(\mathbb{P}^{\pi^*}_{\hat{P}}, \mathbb{P}^{\pi^*}_{P^*})$$

where

$$C(\pi, \pi^*) = \sup_{t \in [H]} \sup_{(s,a) \in \mathcal{S} \times \mathcal{A} : d^{\pi^*,t}_{P^*}(s,a) > 0} \frac{d^{\pi,t}_{P^*}(s,a)}{d^{\pi^*,t}_{P^*}(s,a)}$$

Proof. We derive the proof in three steps:

Step 1.

$$H^2(\mathbb{P}^{\pi}_{\hat{P}}, \mathbb{P}^{\pi}_{P^*}) \leq \sum_{t \in [H-1]} \mathbb{E}_{(s_t,a_t) \sim d^{\pi,t}_{P^*}} \left[H^2 \left(\hat{P}(\cdot|s_t,a_t), P^*(\cdot|s_t,a_t) \right) \right]$$

Step 2.

$$\mathbb{E}_{(s_t,a_t) \sim d^{\pi,t}_{P^*}} \left[H^2 \left(\hat{P}(\cdot|s_t,a_t), P^*(\cdot|s_t,a_t) \right) \right] \leq C(\pi, \pi^*) \cdot \mathbb{E}_{(s_t,a_t) \sim d^{\pi^*,t}_{P^*}} \left[H^2 \left(\hat{P}(\cdot|s_t,a_t), P^*(\cdot|s_t,a_t) \right) \right]$$

Step 3.

$$H^2(\mathbb{P}^{\pi^*}_{\hat{P}}, \mathbb{P}^{\pi^*}_{P^*}) \geq \frac{1}{H} \cdot \mathbb{E}_{(s_t,a_t) \sim d^{\pi^*,t}_{P^*}} \left[H^2 \left(\hat{P}(\cdot|s_t,a_t), P^*(\cdot|s_t,a_t) \right) \right]$$

Proof Step 1:

$$\begin{aligned} H^2(\mathbb{P}^{\pi}_{\hat{P}}, \mathbb{P}^{\pi}_{P^*}) &= 1 - \int_{\mathcal{T}} 1 - \mu_0(s_0) \prod_{t=0}^{H-1} \pi(a_t|s_t) \sqrt{\hat{P}(s_{t+1}|a_t,s_t) P^*(s_{t+1}|s_t,a_t)} d\tau \\ &= 1 - \int_{\mathcal{T}} p^{\pi}_{P^*}(\tau) \cdot \frac{\mu_0(s_0) \prod_{t=0}^{H-1} \pi(a_t|s_t) \sqrt{\hat{P}(s_{t+1}|a_t,s_t) P^*(s_{t+1}|s_t,a_t)}}{\mu_0(s_0) \prod_{t=0}^{H-1} \pi(a_t|s_t) P^*(s_{t+1}|s_t,a_t)} d\tau \\ &= 1 - \int_{\mathcal{T}} p^{\pi}_{P^*}(\tau) \cdot \prod_{t=0}^{H-1} \sqrt{\frac{\hat{P}(s_{t+1}|s_t,a_t)}{P^*(s_{t+1},s_t,a_t)}} d\tau \end{aligned}$$

Next define for ease of notation :

$$\begin{aligned} \alpha_t(s_{t+1}, a_t, s_t) &:= \sqrt{\frac{\hat{P}(s_{t+1}|s_t,a_t)}{P^*(s_{t+1},s_t,a_t)}} \\ \gamma_t(s_t,a_t) &:= \int_{s'} \sqrt{\hat{P}(s_{t+1}|s_t,a_t) P^*(s_{t+1},s_t,a_t)} ds' = \int_{s'} P^*(s'|s_t,a_t) \cdot \alpha_t(s', s_t, a_t) ds' \end{aligned}$$

Notice that γ_t is a BC coefficient i.e

$$1 - \gamma_t(s_t,a_t) = H^2\big(\hat{P}(\cdot|s_t,a_t), P^*(\cdot|s_t,a_t)\big)$$

Using notation above:

$$H^2(\mathbb{P}^\pi_{\hat{P}}, \mathbb{P}^\pi_{P^*}) = 1 - \mathbb{E}_{\tau \sim \mathbb{P}^\pi_{P^*}} \left[\prod_{t=0}^{H-1} \alpha_t(s_{t+1}, s_t, a_t) \right]$$

Using conditional expectation (law of iterated expectation) we change the distribution in the expectation from $\mathbb{P}^\pi_{P^*}$ to the so called state-action distribution $d^{\pi,t}_{P^*}$. To show this argument we show it for state action pair (a_0, s_0, s_1). The rest follows by using the same idea:

$$\begin{aligned}
\mathbb{E}_{\tau \sim \mathbb{P}^\pi_{P^*}} \left[\prod_{t=0}^{H-1} \alpha_t(s_{t+1}, s_t, a_t) \right] &= \mathbb{E}_{s_0,a_0} \left[\mathbb{E}_{s_1|s_0,a_0} \left[\alpha_0(s_1, s_0, a_0) \right] \cdot \mathbb{E}_{a_1 \cup \tau_{[2:H-1]}|s_1} \left[\prod_{t=1}^{H-1} \alpha_t(s_{t+1}, s_t, a_t) \right] \right] \\
&= \int_{s_0,a_0} \mu_0(s_0) \cdot \pi_1(a_0|s_0) \int_{s_1} P^*(s_1|s_0, a_0) \cdot \alpha_0(s_1, s_0, a_0) \\
&\qquad \cdot \mathbb{E}_{a_1 \cup \tau_{[2:H-1]}|s_1} \left[\prod_{t=1}^{H-1} \alpha_t(s_{t+1}, s_t, a_t) \right] \\
&= \int_{s_0,a_0} \mu_0(s_0) \cdot \pi_1(a_0|s_0) \cdot \gamma_0(s_0, a_0) \cdot \Big[\text{,,,,} \Big]
\end{aligned}$$

Notice that

$$\mu_0(s_0) \cdot \pi_1(a_0|s_0) = \int p^\pi_{P^*} d\tau_{[1:H-1]} =: d^{\pi,0}_{P^*}(s_0, a_0) \quad \text{Marginal over } (s_0, a_0)$$

Hence using a recursiv argument we have

$$\begin{aligned}
H^2(\mathbb{P}^\pi_{\hat{P}}, \mathbb{P}^\pi_{P^*}) &= 1 - \mathbb{E}_{\tau \sim \mathbb{P}^\pi_{P^*}} \left[\prod_{t=0}^{H-1} \alpha_t(s_{t+1}, s_t, a_t) \right] \\
&= 1 - \prod_{t=0}^{H-1} \mathbb{E}_{d^{\pi,t}_{P^*}} \Big[\gamma_t(s_t, a_t) \Big]
\end{aligned}$$

Usinge the fact that

$$1 - \prod_i x_i \le \sum_i (1 - x_i) \quad \forall x_i \in [0, 1]$$

and by the fact that $\gamma_t \in [0, 1] \forall t$ we have

$$\begin{aligned}
H^2(\mathbb{P}^\pi_{\hat{P}}, \mathbb{P}^\pi_{P^*}) &\le \sum_{t=0}^{H-1} \mathbb{E}_{d^{\pi,t}_{P^*}} (1 - \gamma_t(s_t, a_t)) \\
&= \sum_{t=0}^{H-1} \mathbb{E}_{d^{\pi,t}_{P^*}} \Big[H^2\big(\hat{P}(\cdot|s_t, a_t), P^*(\cdot|s_t, a_t)\big) \Big]
\end{aligned}$$

Proof Step 2:

$$\begin{aligned}
\mathbb{E}_{(s_t,a_t) \sim d^{\pi,t}_{P^*}} [H^2(\hat{P}(\cdot|s_t, a_t), P^*(\cdot|s_t, a_t))] &= \sum_{s,a} d^{\pi,t}_{P^*}(s, a) \cdot H^2(\hat{P}(\cdot|s, a), P^*(\cdot|s, a)) \\
&= \sum_{s,a} \frac{d^{\pi,t}_{P^*}(s, a)}{d^{\pi^*,t}_{P^*}(s, a)} \cdot d^{\pi^*,t}_{P^*}(s, a) \cdot H^2(\hat{P}(\cdot|s, a), P^*(\cdot|s, a)) \\
&= \sum_{s,a} \frac{d^{\pi,t}_{P^*}(s, a)}{d^{\pi^*,t}_{P^*}(s, a)} \cdot d^{\pi^*,t}_{P^*}(s, a) \cdot H^2(\hat{P}(\cdot|s, a), P^*(\cdot|s, a))
\end{aligned}$$

By definition of the concentrability coefficient:

$$C(\pi, \pi^*) = \sup_{t \in [H]} \sup_{(s,a) \in \mathcal{S} \times \mathcal{A} : d_{P^*}^{\pi^*,t}(s,a) > 0} \frac{d_{P^*}^{\pi,t}(s,a)}{d_{P^*}^{\pi^*,t}(s,a)}$$

Therefore:

$$\frac{d_{P^*}^{\pi,t}(s,a)}{d_{P^*}^{\pi^*,t}(s,a)} \leq C(\pi, \pi^*) \quad \forall t \quad \forall (s,a) \text{ where } d_{P^*}^{\pi^*,t} \geq 0$$

Proof Step 3:
Starting from our previous expression:

$$H^2(\mathbb{P}_{\hat{P}}^{\pi^*}, \mathbb{P}_{P^*}^{\pi^*}) = 1 - \prod_{t=0}^{H-1} \mathbb{E}_{d_{P^*}^{\pi^*,t}}[\gamma_t(s_t, a_t)]$$

Let's denote

$$x_i := 1 - \gamma_i(s_i, a_i) = H^2(\hat{P}(\cdot|s_i, a_i), P^*(\cdot|s_i, a_i))$$

Using the fact that $(1-x) \leq e^{-x}$ for all x, we have:

$$\prod_{i=1}^{H}(1 - x_i) \leq \exp\left(-\sum_{i=1}^{H} x_i\right)$$

By the second-order Taylor expansion of the exponential function:

$$\exp\left(-\sum_{i=1}^{H} x_i\right) \leq 1 - \sum_{i=1}^{H} x_i + \frac{1}{2}\left(\sum_{i=1}^{H} x_i\right)^2$$

Since $x_i \leq 1$ for all i, we know that $\sum_{i=1}^{H} x_i \leq H$, which gives us:

$$\frac{1}{2}\left(\sum_{i=1}^{H} x_i\right)^2 \leq \frac{H}{2}\sum_{i=1}^{H} x_i$$

Therefore:

$$\begin{aligned} H^2(\mathbb{P}_{\hat{P}}^{\pi^*}, \mathbb{P}_{P^*}^{\pi^*}) &\geq \sum_{i=1}^{H} x_i - \frac{1}{2}\left(\sum_{i=1}^{H} x_i\right)^2 \\ &\geq \sum_{i=1}^{H} x_i - \frac{H}{2}\sum_{i=1}^{H} x_i \\ &= \left(1 - \frac{H}{2}\right)\sum_{i=1}^{H} x_i \end{aligned}$$

For the bound $H^2(\mathbb{P}_{\hat{P}}^{\pi^*}, \mathbb{P}_{P^*}^{\pi^*}) \geq \frac{1}{H}\sum_{i=1}^{H} x_i$ to hold, we need:

$$\left(1 - \frac{H}{2}\right)\sum_{i=1}^{H} x_i \geq \frac{1}{H}\sum_{i=1}^{H} x_i$$

This is satisfied when:

$$\sum_{i=1}^{H} x_i \leq \frac{2(H-1)}{H}$$

This condition is typically met for good estimators where Hellinger distances are small. For large H, the bound approaches 2.

Under this condition, we can establish:

$$H^2(\mathbb{P}_{\hat{P}}^{\pi^*}, \mathbb{P}_{P^*}^{\pi^*}) \geq \frac{1}{H} \sum_{t=0}^{H-1} \mathbb{E}_{(s_t,a_t)\sim d_{P^*}^{\pi^*,t}}[H^2(\hat{P}(\cdot|s_t,a_t), P^*(\cdot|s_t,a_t))]$$

□

Lemma 27 (Policy Density Confidence Set)**.** *Assume the following events hold:*

$$E_1 := \left\{H^2(\mathbb{P}_{P^*}^{\hat{\pi}}, \mathbb{P}_{P^*}^{\pi^*}) \leq R_1(\delta_1)\right\}, \quad \textit{such that} \quad P(E_1) \geq 1 - \delta_1,$$
$$E_2 := \left\{H^2(\mathbb{P}_{\hat{P}}^{\pi^*}, \mathbb{P}_{P^*}^{\pi^*}) \leq R_2(\delta_2)\right\}, \quad \textit{such that} \quad P(E_2) \geq 1 - \delta_2,$$

where $\hat{\pi}$ and $\hat{P}$ are estimators of the policy and the transition dynamics, respectively.

Then, under Assumption 4.2, the policy set

$$C_{1-\delta} := \left\{\pi : \sqrt{H^2(\mathbb{P}_{\hat{P}}^{\pi}, \mathbb{P}_{\hat{P}}^{\hat{\pi}})} \leq \sqrt{R_1} + \sqrt{R_2} \cdot \left(1 + \sqrt{\left(1 + \frac{2\sqrt{R_1}}{\gamma_{min}}\right) \cdot H}\right)\right\}$$

is a confidence set of level $1 - \delta = 1 - (\delta_1 + \delta_2)$, i.e.,

$$P(\pi^* \in \Pi_{1-\delta}^{\textit{offline}}) \geq 1 - (\delta_1 + \delta_2).$$

Proof. For ease of notation, let us define:

$$\sqrt{H^2(\mathbb{P}_{P_1}^{\pi_1}, \mathbb{P}_{P_2}^{\pi_2})} =: \|(\pi_1, P_1) - (\pi_2, P_2)\|$$
$$\text{with} \quad \|\cdot\| := \|\cdot\|_{L_2(\mu(\mathbb{R}))}$$

We can then decompose by the triangle inequality:

$$\|(\pi^*, \hat{P}) - (\hat{\pi}, \hat{P})\| \leq \|(\pi^*, \hat{P}) - (\pi^*, P^*)\| + \|(\pi^*, P^*) - (\hat{\pi}, P^*)\| + \|(\hat{\pi}, P^*) - (\hat{\pi}, \hat{P})\|$$

From Lemma 26, we have:

$$\|(\hat{\pi}, P^*) - (\hat{\pi}, \hat{P})\| \leq \sqrt{C(\hat{\pi}, \pi^*) \cdot H} \cdot \|(\pi^*, \hat{P}) - (\pi^*, P^*)\|$$

From Assumption 4.2 and our concentrability coefficient bound, we have:

$$C(\hat{\pi}, \pi^*) \leq 1 + \frac{2\sqrt{R_1}}{\gamma_{min}}$$

From event E_1, we have:

$$\|(\pi^*, P^*) - (\hat{\pi}, P^*)\| \leq \sqrt{R_1}$$

From event E_2, we have:

$$\|(\pi^*, \hat{P}) - (\pi^*, P^*)\| \leq \sqrt{R_2}$$

Then, assuming events $E_1 \cap E_2$ hold jointly, with probability at least $1 - (\delta_1 + \delta_2)$, we have:

$$\|(\pi^*, \hat{P}) - (\hat{\pi}, \hat{P})\| \leq \sqrt{R_1} + \sqrt{R_2} \cdot \left(1 + \sqrt{\left(1 + \frac{2\sqrt{R_1}}{\gamma_{min}}\right) \cdot H}\right)$$

Hence, by construction, the set:

$$\Pi^{\text{offline}}_{1-\delta}(\Pi) := \left\{\pi \in \Pi : \|(\pi, \hat{P}) - (\hat{\pi}, \hat{P})\| \leq \sqrt{R_1} + \sqrt{R_2} \cdot \left(1 + \sqrt{\left(1 + \frac{2\sqrt{R_1}}{\gamma_{min}}\right) \cdot H}\right)\right\}$$

contains π^* with probability at least $1 - (\delta_1 + \delta_2)$. □

B.4 Performance Guarantees

We apply our method of constructing confidence sets based on distributional guarantees for maximum likelihood density estimation to the tabular reinforcement learning setting with state space $\mathcal{S}$ and action space $\mathcal{A}$. We consider deterministic stationary tabular policies ($\Pi = \Pi_S^D$) and stochastic stationary tabular transitions, though the method is versatile to other settings with appropriate adaptation of the corresponding covering numbers (cf. Definitions 17 and 20).

Let $\hat{\pi}$ be the log-loss BC estimator (Equation (1)) of the true policy π^*, and $\hat{P}$ be the MLE estimator (Equation (2)) of the true transition model P^*. The concentration bounds for these estimators are, with probability at least $1 - \delta_1$ and $1 - \delta_2$ respectively:

$$H^2(\mathbb{P}^{\hat{\pi}}_{P^*}, \mathbb{P}^{\pi^*}_{P^*}) \leq R_1 = \frac{4 \cdot |\mathcal{S}| \cdot \log(|\mathcal{A}| \cdot \delta_1^{-1})}{n} \qquad \text{(Corollary 19)}$$

$$H^2(\mathbb{P}^{\pi^*}_{\hat{P}}, \mathbb{P}^{\pi^*}_{P^*}) \leq R_2 = \frac{4 \cdot |\mathcal{S}|^2 \cdot |\mathcal{A}| \cdot \log(nH\delta_2^{-1})}{n} \qquad \text{(Corollary 23)}$$

Additionally, we make the following assumption about the minimum visitation probability under the optimal policy:

Assumption B.3 (Minimum Visitation Probability)**.** *There exists a constant $\gamma_{min} > 0$ such that for all state-action-time tuples with non-zero probability under the optimal policy:*

$$\min_{(s,a,t): d^{\pi^*,t}_{P^*}(s,a) > 0} d^{\pi^*,t}_{P^*}(s,a) \geq \gamma_{min}$$

Under this assumption, the concentrability coefficient is bounded by:

$$C(\hat{\pi}, \pi^*) \leq 1 + \frac{2 \cdot \sqrt{R_1}}{\gamma_{min}}$$

Theorem 28 (Policy Confidence Set)**.** *Under the setting described above and Assumption B.3, with $\delta_1 = \delta_2 = \delta/2$ and defining*

$$\alpha := \sqrt{4 \cdot |\mathcal{S}| \cdot \log(|\mathcal{A}| \cdot 2/\delta)}$$
$$\beta := \sqrt{4 \cdot |\mathcal{S}|^2 \cdot |\mathcal{A}| \cdot \log(nH \cdot 2/\delta)}$$

The policy set

$$\Pi^{\textit{offline}}_{1-\delta} := \left\{\pi : \sqrt{H^2(\mathbb{P}^{\pi}_{\hat{P}}, \mathbb{P}^{\hat{\pi}}_{\hat{P}})} \leq \sqrt{R_1} + \sqrt{R_2} \cdot \left(1 + \sqrt{\left(1 + \frac{2\sqrt{R_1}}{\gamma_{min}}\right) \cdot H}\right)\right\}$$

is a confidence set of level $1-\delta$ containing π^* with probability at least $1-\delta$. The radius of this confidence set is explicitly:

$$\textit{Radius} = \frac{\alpha}{\sqrt{n}} + \frac{\beta}{\sqrt{n}} \cdot \left(1 + \sqrt{H \cdot \left(1 + \frac{2\alpha}{\gamma_{min} \cdot \sqrt{n}}\right)}\right)$$

Proof. The proof follows directly from Lemma 27 by applying our bounds on $H^2(\mathbb{P}^{\hat{\pi}}_{P^*}, \mathbb{P}^{\pi^*}_{P^*})$ and $H^2(\mathbb{P}^{\pi^*}_{\hat{P}}, \mathbb{P}^{\pi^*}_{P^*})$, along with our bound on the concentrability coefficient from Assumption B.3. Setting $\delta_1 = \delta_2 = \delta/2$ and substituting the appropriate values gives us the result. □

C Online Estimation

The underlying setting is described in the Section Problem Setup

C.1 Elliptical Confidence Set

For completeness and to make our paper self-contained, we provide a brief overview of the online preference-based learning approach used in our method. The formulation presented in this section closely follows the work of Saha et al. (2023) and Faury et al. (2020), with adaptations to our specific setting. We include this background to help the reader understand the elliptical confidence set construction that forms a foundation for our theoretical analysis.

In the logistic model, a natural way of computing an estimator $\mathbf{w}_t$ of $\mathbf{w}^*$ given trajectory pairs $\{(\tau^1_\ell, \tau^2_\ell)\}_{\ell=1}^{t-1}$ and preference feedback values $\{o_\ell\}_{\ell=1}^{t-1}$ is via maximum likelihood estimation. At time t the regularized log-likelihood (or negative cross-entropy loss) of a parameter $\mathbf{w}$ can be written as:

$$\begin{aligned}\mathcal{L}_t^\lambda(\mathbf{w}) = \sum_{\ell=1}^{t-1} \Big(& o_\ell \log(\sigma(\langle \phi(\tau^1_\ell) - \phi(\tau^2_\ell) \rangle, \mathbf{w} \rangle)) - \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \\ & + (1-o_\ell)\log\left(1 - \sigma(\langle \phi(\tau^1_\ell) - \phi(\tau^2_\ell), \mathbf{w}\rangle)\right)\Big),\end{aligned}$$

where $\lambda > 0$ is a regularization parameter. The function $\mathcal{L}_t^\lambda$ is strictly concave for $\lambda > 0$. The maximum likelihood estimator $\hat{\mathbf{w}}_t^{\text{MLE}}$ can be written as $\hat{\mathbf{w}}_t^{\text{MLE}} = \arg\max_{\mathbf{w}\in\mathbb{R}^d} \mathcal{L}_t^\lambda(\mathbf{w})$. Unfortunately, $\hat{\mathbf{w}}_t^{\text{MLE}}$ may not satisfy the boundedness Assumption 1, so we instead make use of a projected version of $\hat{\mathbf{w}}_t^{\text{MLE}}$. Following Faury et al. (2020), and recalling Assumption 1, we define a data matrix and a transformation of $\hat{\mathbf{w}}_t^{\text{MLE}}$ given by

$$\mathbf{V}_t = \kappa\lambda\mathbf{I}_d + \sum_{\ell=1}^{t-1} \left(\phi(\tau^1_\ell) - \phi(\tau^2_\ell)\right)\left(\phi(\tau^1_\ell) - \phi(\tau^2_\ell)\right)^\top$$

$$g_t(\mathbf{w}) = \sum_{\ell=1}^{t-1} \sigma(\langle \phi(\tau^1_\ell) - \phi(\tau^2_\ell), \mathbf{w}\rangle)\left(\phi(\tau^1_\ell) - \phi(\tau^2_\ell)\right) + \lambda\mathbf{w}$$

Then, the projected parameter, along with its confidence set, is given by

$$\mathbf{w}_t^{Proj} = \underset{\mathbf{w} \text{ s.t. } \|\mathbf{w}\|\le W}{\arg\min} \; \|g_t(\mathbf{w}) - g_t(\hat{\mathbf{w}}_t^{\text{MLE}})\|_{\mathbf{V}_t^{-1}}$$

$$\mathcal{C}_t(\delta) = \{\mathbf{w} \text{ s.t. } \|\mathbf{w} - \mathbf{w}_t^P\|_{\mathbf{V}_t} \le 2\kappa\beta_t(\delta)\}$$

where $\beta_t(\delta) = \sqrt{\lambda}W + \sqrt{\log(1/\delta) + 2d\log\left(1 + \frac{tB^2}{\kappa\lambda d}\right)}$. We restate a bound by Faury et al. (2020) that shows the probability of $\mathbf{w}_*$ being in $\mathcal{C}_t(\delta)$ for all $t \ge 1$ can be lower bounded.

Lemma 29 (Confidence Set Coverage)**.** *Let $\delta \in (0, 1]$ and define the event that $\mathbf{w}_*$ is in the confidence interval $\mathcal{C}_t(\delta)$ for all $t \in \mathbb{N}$:*

$$E_{w^*} = \{\forall t \geq 1, \mathbf{w}_* \in \mathcal{C}_t(\delta)\}.$$

Then $\mathbb{P}(E_{w^*}) \geq 1 - \delta$.

Proof. This follows from Faury et al. (2020) with a slight modification to account for our bounded feature assumption. □

This elliptical confidence set construction, which has its roots in generalized linear bandits (Filippi et al., 2010; Faury et al., 2020), forms a critical component of our online learning algorithm. By maintaining and updating these confidence sets as new preference data is collected, our algorithm can efficiently balance exploration and exploitation to identify the optimal policy. The confidence bounds ensure that with high probability, the true reward parameter lies within our constructed set throughout the learning process, which is essential for the regret guarantees we derive in the following sections.

C.2 Norm Relation Between Data Matrices

For completeness, we restate key results from Saha et al. (2023) concerning the relationships between various data matrices that arise in our analysis. These results are included to ensure the appendix is self-contained and to provide context for our subsequent analysis. The full proofs of these results can be found in the original paper.

Saha et al. (2023) establishes relationships between three key matrices:

- V_t - The empirical data matrix constructed from observed trajectories
- $\overline{V}_t^{P^*}$ - The expected data matrix under the true transition dynamics P^*
- $\overline{V}_t$ - The expected data matrix under the estimated transition dynamics $\hat{P}_t$

These matrices are defined as follows:

$$\begin{aligned} V_t &= \kappa\lambda \mathbf{I}_d + \sum_{\ell=1}^{t-1} \big(\phi(\tau_\ell^1) - \phi(\tau_\ell^2)\big)\big(\phi(\tau_\ell^1) - \phi(\tau_\ell^2)\big)^\top \\ \overline{V}_t^{P^*} &= \kappa\lambda \mathbf{I}_d + \sum_{\ell=1}^{t-1} \big(\phi(\pi_\ell^1) - \phi(\pi_\ell^2)\big)\big(\phi(\pi_\ell^1) - \phi(\pi_\ell^2)\big)^\top \\ \overline{V}_t &= \kappa\lambda \mathbf{I}_d + \sum_{\ell=1}^{t-1} \big(\phi^{\hat{P}_\ell}(\pi_\ell^1) - \phi^{\hat{P}_\ell}(\pi_\ell^2)\big)\big(\phi^{\hat{P}_\ell}(\pi_\ell^1) - \phi^{\hat{P}_\ell}(\pi_\ell^2)\big)^\top \end{aligned}$$

Where $\phi(\pi)$ represents the expected feature vector under policy π and the true transition dynamics P^*, while $\phi^{\hat{P}_t}(\pi)$ represents the expected feature vector under policy π and the estimated transition dynamics $\hat{P}_t$.

Saha et al. (2023) introduces a precision event that relates the empirical matrix V_T to the expected matrix $\overline{V}_T^{P^*}$:

$$E_{\overline{V}_T^{P^*}} = \{\overline{V}_T^{P^*} \preceq 2V_T + 84B^2 d \log((1 + 2T)/\delta)\mathbf{I}_d\}$$

Under this event, they establish the following bound:

Lemma 30 (Adapted from Saha et al. (2023) Corollary 1)**.** *Under Assumption 1, conditioned on event $E_{w^*} \cap E_{\overline{V}_T^{P^*}}$, for any $t \in [T]$*

$$\|\mathbf{w}^* - \mathbf{w}_t^L\|_{\overline{V}_t^{P^*}} \leq 4\kappa\beta_t(\delta) + \alpha_{d,T}(\delta),$$

where $\alpha_{d,T}(\delta) = 20BW\sqrt{d\log(T(1+2T)/\delta)}$. Furthermore, if $\delta \leq 1/e$, then $\mathbb{P}(E_{w^*} \cap E_{\overline{V}_T^{P^*}}) \geq 1 - \delta - \delta\log_2 T$.

Additionally, Saha et al. (2023) relates norms based on the matrix $\overline{V}_t^{P^*}$ with those based on $\overline{V}_t$:

Lemma 31 (Adapted from Saha et al. (2023) Lemma 3)**.** *Let $\overline{\mathcal{E}}_0$ be the event that for all $t \in \mathbb{N}$,*

$$\|\mathbf{w}_t^{proj} - \mathbf{w}_*\|_{\overline{V}_t} \leq \sqrt{2}\|\mathbf{w}_t^{proj} - \mathbf{w}_*\|_{\overline{V}_t^{P^*}} + \sqrt{\sum_{\ell=1}^{t-1} 4\left(\hat{B}_\ell\left(\pi, 2WB, \frac{\delta'}{8\ell^3|A|^{|S|}}\right)\right)^2 + \frac{1}{t}}$$

where $\delta' = \frac{\delta}{(1+4W)^d}$ and $\epsilon = \frac{1}{2\kappa\lambda+4B^2t^\alpha}$. Then $\mathbb{P}\left(\overline{\mathcal{E}}_0\right) \geq 1-\delta$.

Note that the bonus function $\hat{B}$ is defined in Lemma 32
These norm relations from Saha et al. (2023) are essential in our regret analysis, as they allow us to relate confidence bounds across different probability spaces and to bound the regret of our algorithm.

C.3 Transition Estimation and Bonus Terms

Note that the offline estimator of the transition probabilities based on the log-loss MLE in Equation (2), when the state-action space is discrete, is equivalent to the following count-based estimator (derivable using a simple Lagrange multiplier argument):

$$\hat{P}_{\text{offline}}(s'|s,a) = \frac{N_{\text{offline}}(s'|s,a)}{N_{\text{offline}}(s,a)}$$

where

$$N_{\text{offline}}(s,a) := \sum_{i\in[n]}\sum_{h\in[H]} \mathbb{I}\{s_h^i = s, a_h^i = a\}$$
$$N_{\text{offline}}(s'|s,a) := \sum_{i\in[n]}\sum_{h\in[H]} \mathbb{I}\{s_{h+1}^i = s', s_h^i = s, a_h^i = a\}$$

This equivalence allows us to initialize the online estimation process with the count estimator from the offline data (see line 3 in Algorithm 1), yielding the combined estimator for the transition model:

$$\hat{P}_t(s'|s,a) := \frac{N_{\text{offline}}(s'|s,a) + N_t(s'|s,a)}{N_{\text{offline}}(s,a) + N_t(s,a)} \tag{9}$$

From this estimator, we adapt two key lemmas from Chatterji et al. (2021) that will define our notion of bonus terms.

Lemma 32 (Moment Transition Difference Error)**.** *Consider the transition count estimator $\hat{P}_t$ from Equation (9). Further assume the trajectory data follows a martingale structure adapted to the natural filtration of the problem. For any fixed policy $\pi \in \Pi$ and any scalar function $f : \mathcal{T} \to \mathbb{R}$ such that $|f(\tau)| < \eta$, with probability at least $1-\delta$ for all $t \in \mathbb{N}$:*

$$\mathbb{E}_{\mathbb{P}_{P^*}^{\pi}}[f(\tau)] - \mathbb{E}_{\mathbb{P}_{\hat{P}_t}^{\pi}}[f(\tau)] \leq \mathbb{E}_{\mathbb{P}_{\hat{P}_t}^{\pi}}\left[\sum_{h\in[H]} \xi_{s_h,a_h}^t(\eta,\delta)\right] =: \hat{B}_t(\pi,\eta,\delta)$$

where

$$\xi^t_{s_h,a_h}(\eta,\delta) := \min\left(2\eta, 4\eta\sqrt{\frac{H\log(|\mathcal{S}|\cdot|\mathcal{A}|)+\log\left(\frac{6\log(N_t(s_h,a_h)+N_{\textit{offline}}(s_h,a_h))}{\delta}\right)}{N_t(s_h,a_h)+N_{\textit{offline}}(s_h,a_h)}}\right)$$

The term $\hat{B}_t(\pi,\eta,\delta)$ serves as our "bonus" term.

Proof. Our combined estimator incorporates both online data (adapted to the natural filtration) and offline data (assumed i.i.d.). We can artificially treat the offline data as though it were adapted to the natural filtration as well, by considering it as "past" observations. This allows us to directly apply the proof methodology from Chatterji et al. (2021) (Lemma B.1) to our combined count estimator.

The key insight is that the martingale structure of the estimation error is preserved when combining offline and online counts, with the benefit of reduced variance due to the increased denominator $(N_t(s_h,a_h)+N_{\text{offline}}(s_h,a_h))$. This directly translates to tighter confidence bounds compared to using only online data. □

We now present a stronger version of the lemma that holds uniformly for all policies π.

Lemma 33 (Uniform Moment Transition Difference Error)**.** *Consider the transition count estimator $\hat{P}_t$ from Equation (9). Further assume the trajectory data follows a martingale structure adapted to the natural filtration of the problem. For any scalar function $f:\mathcal{T}\to\mathbb{R}$ such that $|f(\tau)|<\eta$ and for any $\epsilon>0$, with probability at least $1-\delta$ for all $t\in\mathbb{N}$ and all $\pi\in\Pi$:*

$$\mathbb{E}_{\mathbb{P}^{\pi}_{\hat{P}_t}}[f(\tau)] - \mathbb{E}_{\mathbb{P}^{\pi}_{P^*}}[f(\tau)] \le \underbrace{\mathbb{E}_{\mathbb{P}^{\pi}_{P^*}}\left[\sum_{h\in[H]}\overline{\xi}^t_{s_h,a_h}(\eta,\delta,\epsilon)\right]}_{=:B_t(\pi,\eta,\delta,\epsilon)} + \epsilon$$

where

$$\overline{\xi}^t_{s_h,a_h}(\eta,\delta,\epsilon) := \min\left(2\eta, 4\eta\sqrt{\frac{H\log(|\mathcal{S}|\cdot|\mathcal{A}|)+|\mathcal{S}|\log\left(\left\lceil\frac{4\eta H}{\epsilon}\right\rceil\right)+\log\left(\frac{6\log(N_t(s_h,a_h)+N_{\textit{offline}}(s_h,a_h))}{\delta}\right)}{N_t(s_h,a_h)+N_{\textit{offline}}(s_h,a_h)}}\right)$$

Proof. The proof follows by applying similar techniques as in Lemma 32, but with additional care to ensure uniformity across all policies.

As before, we can artificially treat the offline data as adapted to the natural filtration. The uniform convergence over the policy class Π is achieved by applying a covering argument and the union bound, following the methodology in Chatterji et al. (2021) (Lemma B.2). The additional term $|\mathcal{S}|\log\left(\left\lceil\frac{4\eta H}{\epsilon}\right\rceil\right)$ appears due to this covering, which introduces an ϵ-discretization of the policy space.

The combined offline and online counts in the denominator $(N_t(s_h,a_h)+N_{\text{offline}}(s_h,a_h))$ provide tighter uniform confidence bounds compared to using online data alone. □

To provide further intuition, we elaborate on the meaning and significance of the terms $\hat{B}_t$ and B_t introduced in the previous lemmas. In reinforcement learning literature, these would be referred to as the "empirical bonus" and "true bonus," respectively. Both terms quantify the concentration of our estimators around their true values.

The empirical bonus $\hat{B}_t(\pi,\eta,\delta)$ represents the expected sum of state-action-level uncertainty terms $\xi^t_{s_h,a_h}(\eta,\delta)$ under the *estimated* transition model $\hat{P}_t$. Importantly, this term can be directly computed from observed data.

In contrast, the true bonus $B_t(\pi, \eta, \delta, \epsilon)$ represents the expected sum of uncertainty terms $\overline{\xi}^t_{s_h,a_h}(\eta, \delta, \epsilon)$ under the *true* transition model P^*. This term cannot be directly computed as it depends on the unknown true model.

For our regret analysis, we need to relate these two quantities. The following lemma provides a crucial connection, showing that the empirical bonus $\hat{B}_t$ can be bounded in terms of the true bonus B_t uniformly across all policies π.

Lemma 34 (Relationship Between Empirical and True Bonus Terms)**.** *Let $\eta, \epsilon > 0$. For all policies $\pi \in \Pi$ simultaneously and for all $t \in \mathbb{N}$, with probability at least $1 - \delta$:*

$$\hat{B}_t(\pi, \eta, \delta) \leq 2B_t(\pi, 2H\eta, \delta, \epsilon) + \epsilon$$

Proof. Define the function $f : \mathcal{T} \to \mathbb{R}$ as:

$$f(\tau) := \sum_{h \in [H]} \xi^t_{s_h,a_h}(\eta, \delta)$$

By construction, $\hat{B}_t(\pi, \eta, \delta) = \mathbb{E}_{\mathbb{P}^\pi_{\hat{P}_t}}[f(\tau)]$. Since $\xi^t_{s_h,a_h}(\eta, \delta) \leq 2\eta$ for all state-action pairs, we have $|f(\tau)| \leq 2\eta H$.

Applying Lemma 33 with this $f(\tau)$ and the bound $2\eta H$:

$$\mathbb{E}_{\mathbb{P}^\pi_{\hat{P}_t}}[f(\tau)] - \mathbb{E}_{\mathbb{P}^\pi_{P^*}}[f(\tau)] \leq \mathbb{E}_{\mathbb{P}^\pi_{P^*}}\left[\sum_{h \in [H]} \overline{\xi}^t_{s_h,a_h}(2\eta H, \delta, \epsilon)\right] + \epsilon$$

By definition, the right-hand side equals $B_t(\pi, 2H\eta, \delta, \epsilon) + \epsilon$. Therefore:

$$\begin{aligned}\hat{B}_t(\pi, \eta, \delta) &= \mathbb{E}_{\mathbb{P}^\pi_{\hat{P}_t}}[f(\tau)] \\ &\leq \mathbb{E}_{\mathbb{P}^\pi_{P^*}}[f(\tau)] + B_t(\pi, 2H\eta, \delta, \epsilon) + \epsilon\end{aligned}$$

From Lemma 32, we know that:

$$\mathbb{E}_{\mathbb{P}^\pi_{P^*}}[f(\tau)] \leq \mathbb{E}_{\mathbb{P}^\pi_{\hat{P}_t}}[f(\tau)] + \hat{B}_t(\pi, \eta, \delta) = \hat{B}_t(\pi, \eta, \delta) + \hat{B}_t(\pi, \eta, \delta) = 2\hat{B}_t(\pi, \eta, \delta)$$

This gives us:

$$\begin{aligned}\hat{B}_t(\pi, \eta, \delta) &\leq 2\hat{B}_t(\pi, \eta, \delta) + B_t(\pi, 2H\eta, \delta, \epsilon) + \epsilon \\ \Rightarrow -\hat{B}_t(\pi, \eta, \delta) &\leq B_t(\pi, 2H\eta, \delta, \epsilon) + \epsilon \\ \Rightarrow \hat{B}_t(\pi, \eta, \delta) &\leq B_t(\pi, 2H\eta, \delta, \epsilon) + \epsilon\end{aligned}$$

Therefore, the lemma statement follows. □

This lemma is instrumental for our regret analysis as it allows us to work with B_t instead of $\hat{B}_t$. The advantage is that B_t involves expectations with respect to the true transition model P^*, which makes it more amenable to theoretical analysis. By establishing this relationship, we effectively account for the transition estimation error and can focus on controlling the difference between empirical and true moments, which is a more tractable problem in our analytical framework.

C.4 Policy Set Π_t and Proof Lemma 7

Recall that we define the policy set Π_t to draw from in line 7 of Algorithm 1 as

$$\begin{aligned}\Pi_t := \big\{\, \pi \in \Pi^{\text{offline}}_{1-\delta} \;\big|\, \forall \pi' \in \Pi^{\text{offline}}_{1-\delta} :&\\ &\big\langle \phi^{\hat{P}_t}(\pi) - \phi^{\hat{P}_t}(\pi'), w^{\text{proj}}_t \big\rangle + \gamma_t \cdot \|\phi^{\hat{P}_t}(\pi) - \phi^{\hat{P}_t}(\pi')\|_{\overline{\mathbf{V}}_t^{-1}}\\ &+ \hat{B}_t(\pi, 2WB, \delta') + \hat{B}_t(\pi', 2WB, \delta') \geq 0 \,\big\}\end{aligned}$$

where $\Pi^{\text{offline}}_{1-\delta}$ is derived in Theorem 4. The radius γ_t is defined as

$$\gamma_t = \sqrt{2}(4\kappa\beta_t(\delta) + \alpha_{d,T}(\delta)) + 2\sqrt{\sum_{\ell=1}^{t-1}\left(\hat{B}_\ell\left(\pi, 2WB, \frac{\delta'}{8\ell^3|A|^{|S|}}\right)\right)^2 + \frac{1}{t}}$$

Then Lemma 7 state that with high probability, $\pi^* \in \Pi_t \quad \forall t \in [T]$

Proof of Lemma 7. We begin by conditioning on the following events:

- $E_{\text{offline}} = \{\pi^* \in \Pi^{\text{offline}}_{1-\delta}\}$ from Theorem 4
- E_{w^*} from Lemma 30 (confidence set for w^*)
- $E_{\overline{V}_T^{P^*}}$ from Lemma 30 (relation for data matrices)
- $\overline{\mathcal{E}}_0$ from Lemma 31 (estimated norm relation)
- $\mathcal{E}_3$ from Lemma 32 (bounds on the bonus terms $\hat{B}_t$)

By the union bound, these five events hold simultaneously with probability at least $1 - 5\delta$.

By the optimality condition, we have:

$$0 \leq \langle \phi(\pi^*) - \phi(\pi'), w^* \rangle = \langle \mathbb{E}_{\mathbb{P}^{\pi^*}_{P^*}} \phi(\tau) - \mathbb{E}_{\mathbb{P}^{\pi'}_{P^*}} \phi(\tau), w^* \rangle$$

Then, by event $\mathcal{E}_3$ and defining $f(\tau) := \langle \phi(\tau), w^* \rangle$, we have from **??** that $|f(\tau)| \leq 2WB$, which yields:

$$\langle \phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi'), w^* \rangle + \hat{B}_t(\pi^*, 2WB, \delta/|\mathcal{A}|^{|\mathcal{S}|}) + \hat{B}_t(\pi', 2WB, \delta/|\mathcal{A}|^{|\mathcal{S}|})$$

where the probability parameter accounts for any $\pi' \in \Pi$, which covers the case of the offline confidence set being the whole policy space (i.e., not having enough offline data for learning).

Next, we bound the term:

$$\begin{aligned}\langle \phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi'), w^* \rangle &= \langle \phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi'), w^{\text{proj}}_t \rangle + \langle \phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi'), w^* - w^{\text{proj}}_t \rangle\\ &\leq \langle \phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi'), w^{\text{proj}}_t \rangle + \|\phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi')\|_{\overline{V}_t^{-1}} \cdot \|w^{\text{proj}}_t - w^*\|_{\overline{V}_t}\end{aligned}$$

We can now use event $\overline{\mathcal{E}}_0$:

$$\|w^{\text{proj}}_t - w^*\|_{\overline{V}_t} \leq \sqrt{2}\|w^{\text{proj}}_t - w_*\|_{\overline{V}_t^{P^*}} + 2\sqrt{\sum_{\ell=1}^{t-1}\left(\hat{B}_\ell\left(\pi, 2WB, \frac{\delta'}{8\ell^3|A|^{|S|}}\right)\right)^2 + \frac{1}{t}}$$

Using events $E_{w^*} \cap E_{\overline{V}_T^{P^*}}$, we get:

$$\|w^{\text{proj}}_t - w^*\|_{\overline{V}_t} \leq \sqrt{2}(4\kappa\beta_t(\delta) + \alpha_{d,T}(\delta)) + 2\sqrt{\sum_{\ell=1}^{t-1}\left(\hat{B}_\ell\left(\pi, 2WB, \frac{\delta'}{8\ell^3|A|^{|S|}}\right)\right)^2 + \frac{1}{t}} =: \gamma_t$$

Putting these results together yields that $\pi^* \in \Pi_t$ for all $t \in \mathbb{N}$ under the event E_{offline}.

The probability of this event is at least $1 - 5\delta$ by the union bound of all the events we conditioned on. By rescaling $\delta \mapsto \delta/5$, we obtain the desired result with probability at least $1 - \delta$. □

C.5 Regret Bound

In this section, we provide a lemma as an intermediate step toward the full proof of the regret analysis of BRIDGE. This lemma separates the upper bound on the regret into three distinct terms, each of which we further analyze in Appendix D.

Lemma 35 (Regret Analysis)**.** *Under the following events:*

- $E_{\textit{offline}} = \{\pi^* \in \Pi_{1-\delta}^{\textit{offline}}\}$ *from Theorem 4*
- E_{w^*} *from Lemma 30 (confidence set for* w^**)*
- $E_{\overline{V}_T^{P^*}}$ *from Lemma 30 (relation for data matrices)*
- $\overline{\mathcal{E}}_0$ *from Lemma 31 (estimated norm relation)*
- $\mathcal{E}_3$ *from Lemma 32 (bounds on the bonus terms* $\hat{B}_t$*)*

the regret of BRIDGE *Algorithm 1 is upper bounded by:*

$$R_T \leq 2 \cdot \underbrace{\gamma_T}_{\textit{Term 1}} \cdot \underbrace{\sqrt{T \sum_{t\in[T]} \|\phi^{\hat{P}_t}(\pi_t^1) - \phi^{\hat{P}_t}(\pi_t^2)\|_{\overline{V}_t^{-1}}}}_{\textit{Term 2}} + \underbrace{\sum_{i\in\{1,2\}} \sum_{t\in[T]} \hat{B}_t(\pi_t^i, 4WB, \delta)}_{\textit{Term 3}}$$

where

$$\gamma_T = \sqrt{2}(4\kappa \cdot \beta_T(\delta) + \alpha_{d,T}(\delta)) + \frac{1}{T} + 4\sqrt{\sum_{i\in\{1,2\}} \sum_{t\in[T]} B_T(\pi_t^i, 4HWB, \delta, \epsilon)^2 + 96T\epsilon HWB}$$

and

- $\alpha_{d,T}(\delta) = 20BW\sqrt{d\log(T(1+2T)/\delta)}$
- $\beta_T(\delta) = \sqrt{\lambda}W + \sqrt{\log(1/\delta) + 2d\log\left(1 + \frac{TB^2}{\kappa\lambda d}\right)}$.

Proof. We start by writing

$$\begin{aligned} 2r_t &= \langle \phi(\pi^*) - \phi(\pi_t^1), w^* \rangle + \langle \phi(\pi^*) - \phi(\pi_t^2), w^* \rangle \\ &= \langle \phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi_t^1), w^* \rangle + \langle \phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi_t^2), w^* \rangle + 2\langle \phi^{\hat{P}_t}(\pi^*) - \phi(\pi^*), w^* \rangle \\ &+ \langle \phi^{\hat{P}_t}(\pi_t^1) - \phi(\pi_t^1), w^* \rangle + \langle \phi^{\hat{P}_t}(\pi_t^2) - \phi(\pi_t^2), w^* \rangle \end{aligned}$$

Then, by Lemma 32, we have with probability at least $1 - \delta$ for each of the following:

$$\begin{aligned} 2\langle \phi(\pi^*) - \phi^{\hat{P}_t}(\pi^*), w^* \rangle &\leq 2\hat{B}_t(\pi^*, 4WB, \delta) \\ \langle \phi^{\hat{P}_t}(\pi_t^1) - \phi(\pi_t^1), w^* \rangle &\leq \hat{B}_t(\pi_t^1, 4WB, \delta) \\ \langle \phi^{\hat{P}_t}(\pi_t^2) - \phi(\pi_t^2), w^* \rangle &\leq \hat{B}_t(\pi_t^2, 4WB, \delta) \end{aligned}$$

By the union bound, with high probability:

$$2r_t \leq \langle \phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi_t^1), w^* \rangle + \langle \phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi_t^2), w^* \rangle + \hat{B}_t(\pi_t^1, 4WB, \delta) + \hat{B}_t(\pi_t^2, 4WB, \delta) + 2\hat{B}_t(\pi^*, 4WB, \delta)$$

Next, we observe that:

$$\begin{aligned}
&\langle \phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi_t^1), w^* \rangle + \langle \phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi_t^2), w^* \rangle \\
&\leq \langle \phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi_t^1), w_t^{\text{proj}} \rangle + \langle \phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi_t^2), w_t^{\text{proj}} \rangle \\
&+ \|w^* - w_t^{\text{proj}}\|_{\overline{V}_t} \left(\|\phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi_t^1)\|_{\overline{V}_t^{-1}} + \|\phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi_t^2)\|_{\overline{V}_t^{-1}} \right)
\end{aligned}$$

Conditioning on the joint event $\mathcal{E}_0 \cap E_{w^*} \cap E_{\overline{V}_t^{P^*}}$, we have with high probability:

$$\begin{aligned}
&\langle \phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi_t^1), w^* \rangle + \langle \phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi_t^2), w^* \rangle \\
&\leq \langle \phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi_t^1), w_t^{\text{proj}} \rangle + \langle \phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi_t^2), w_t^{\text{proj}} \rangle \\
&+ \gamma_t \cdot \left(\|\phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi_t^1)\|_{\overline{V}_t^{-1}} + \|\phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi_t^2)\|_{\overline{V}_t^{-1}} \right)
\end{aligned}$$

Using Lemma 32 again, the following holds with high probability:

$$\begin{aligned}
2\langle \phi(\pi^*) - \phi^{\hat{P}_t}(\pi^*), w_t^{\text{proj}} \rangle &\leq 2\hat{B}_t(\pi^*, 4WB, \delta) \\
\langle \phi^{\hat{P}_t}(\pi_t^1) - \phi(\pi_t^1), w_t^{\text{proj}} \rangle &\leq \hat{B}_t(\pi_t^1, 4WB, \delta) \\
\langle \phi^{\hat{P}_t}(\pi_t^2) - \phi(\pi_t^2), w_t^{\text{proj}} \rangle &\leq \hat{B}_t(\pi_t^2, 4WB, \delta)
\end{aligned}$$

Putting everything together yields:

$$\begin{aligned}
2r_t \leq 2\gamma_t &\left(\|\phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi_t^1)\|_{\overline{V}_t^{-1}} + \|\phi^{\hat{P}_t}(\pi^*) - \phi^{\hat{P}_t}(\pi_t^2)\|_{\overline{V}_t^{-1}} \right) \\
&+ 2\hat{B}_t(\pi_t^1, 4WB, \delta) + 2\hat{B}_t(\pi_t^2, 4WB, \delta) + 4\hat{B}_t(\pi^*, 4WB, \delta)
\end{aligned}$$

Under the event $\pi^* \in \Pi_t$ from Lemma 7 and using the fact that $\pi_t^1, \pi_t^2 \in \Pi_t$, we have:

$$2r_t \leq \gamma_t \|\phi^{\hat{P}_t}(\pi_t^2) - \phi^{\hat{P}_t}(\pi_t^1)\|_{\overline{V}_t^{-1}} + 4\hat{B}_t(\pi_t^1, 4WB, \delta) + 4\hat{B}_t(\pi_t^2, 4WB, \delta)$$

Hence, the regret is:

$$\begin{aligned}
R_T &= \sum_{t \in [T]} 2r_t \\
&\leq \sum_{t \in [T]} \left(\gamma_t \|\phi^{\hat{P}_t}(\pi_t^2) - \phi^{\hat{P}_t}(\pi_t^1)\|_{\overline{V}_t^{-1}} + 4\hat{B}_t(\pi_t^1, 4WB, \delta) + 4\hat{B}_t(\pi_t^2, 4WB, \delta) \right) \\
&\leq \gamma_T \sqrt{T \sum_{t \in [T]} \|\phi^{\hat{P}_t}(\pi_t^2) - \phi^{\hat{P}_t}(\pi_t^1)\|^2_{\overline{V}_t^{-1}}} + \sum_{t \in [T]} \left(4\hat{B}_t(\pi_t^1, 4WB, \delta) + 4\hat{B}_t(\pi_t^2, 4WB, \delta) \right)
\end{aligned}$$

Note that by Lemma 32, with high probability:

$$\hat{B}_t(\pi_l^i, 2WB, \delta)^2 \leq 4B_t(\pi_l^i, 2HWB, \delta, \epsilon) + 24\epsilon HWB$$

Plugging this into γ_t yields:

$$\begin{aligned}
\gamma_t &\leq \sqrt{2}(4\kappa\beta_t(\delta) + \alpha_{d,T}(\delta)) + \frac{1}{t} \\
&+ 4\sqrt{\sum_{\ell=1}^{t-1} B_t^2(\pi_t^1, 4HSB, \delta_t', \epsilon) + B_t^2(\pi_t^2, 4HSB, \delta_t') + 96(t-1)HWB} \quad \forall t
\end{aligned}$$

This completes the proof of the claimed result. □

D Regret Analysis: Theorem 8

In this section, we present the complete regret analysis of our BRIDGE algorithm. We recommend that readers first review Appendix A, where we analyze a simplified setting in which the dynamics are assumed to be known. This simplified case captures the core idea of our approach: constraining the set of policies considered during online preference learning using a confidence interval derived from offline behavioral cloning estimation (see **??**).

The key difference in the present analysis is that we now incorporate the estimation of the transition model. Specifically, we first estimate the transition model offline and then use this estimate as the starting point for online transition estimation. This approach reduces the error due to transition uncertainty by a factor of $\mathcal{O}(1/\sqrt{n})$, which is the same rate of improvement we achieve for the policy estimation through behavioral cloning. As we will show, this allows our algorithm to effectively leverage offline demonstrations to reduce both sources of uncertainty, resulting in substantially improved regret bounds.

Theorem 36 (Regret Bound with Offline-Enhanced Exploration)**.** *Let n be the number of offline demonstrations with minimum visitation probability $\gamma_{\min} > 0$ for state-action pairs. With probability at least $1 - \delta$, the regret of the algorithm is bounded by:*

$$R_T \leq 2 \cdot \underbrace{\gamma_T}_{\textit{Term 1}} \cdot \underbrace{\sqrt{T \cdot \log\left(1 + \frac{\tilde{O}\left(B^2 \cdot H \cdot |S|^2 \cdot \min\left\{\frac{T}{n}, \ln(T)\right\} + \frac{T \cdot |S| \cdot B^2 \cdot \sqrt{|A| \cdot H}}{\sqrt{n \cdot \gamma_{\min}}}\right)}{d}\right)}}_{\textit{Term 2}}$$

$$+ \underbrace{\tilde{O}\left(H|\mathcal{S}|\sqrt{\frac{|\mathcal{A}|TH}{n \cdot \gamma_{\min}}} + \frac{H^{5/2}WB\sqrt{T}}{\sqrt{n \cdot \gamma_{\min}}} + H^2WB \cdot \sqrt{T} \cdot \frac{|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/4}}{n^{1/4}}\right)}_{\textit{Term 3}}$$

where

$$\gamma_T = \tilde{O}\left((\kappa + BW)\sqrt{d\log(T)} + H^2WB|S| \cdot \sqrt{\min\left\{\log(T), \frac{T}{n \cdot \gamma_{\min}}\right\}} + \sqrt{HWB}\right)$$

and we have set $\epsilon = \frac{1}{T}$ to optimize the bound.

From Lemma 35, we analyze the three key terms in our regret bound: the confidence multiplier (Term 1), the logarithmic determinant ratio (Term 2), and the bonus function summation (Term 3). Each term is examined in detail in the following subsections.

$$\text{Term 1} = \gamma_T = \sqrt{2}(4\kappa \cdot \beta_T(\delta) + \alpha_{d,T}(\delta)) + \frac{1}{T} + 4\sqrt{\sum_{i\in\{1,2\}} \sum_{t\in[T]} B_T(\pi_t^i, 4HWB, \delta, \epsilon)^2 + 96T\epsilon HWB}$$

$$\text{Term 2} = \sqrt{T \sum_{t\in[T]} \|\phi^{\hat{P}_t}(\pi_t^1) - \phi^{\hat{P}_t}(\pi_t^2)\|_{\overline{V}_t^{-1}}}$$

$$\text{Term 3} = \sum_{i\in\{1,2\}} \sum_{t\in[T]} \hat{B}_t(\pi_t^i, 4WB, \delta)$$

D.1 Term 1: Asymptotic bound

We derive an asymptotic bound for Term 1 in Theorem 36 via Lemma 37. The auxiliary lemmata used in the proof of Lemma 37 are found in Appendix D.1.1.

Lemma 37. *The asymptotic bound on γ_T can be expressed as:*

$$\gamma_T = \tilde{O}\left((\kappa + BW)\sqrt{d\log(T)} + H^2 WB|S| \cdot \sqrt{\min\left\{\log(T), \frac{T}{n \cdot \gamma_{\min}}\right\}} + \sqrt{T\epsilon HWB}\right)$$

Proof. We analyze each term in the expression for γ_T separately.

Step 1: Analyze $\sqrt{2}(4\kappa \cdot \beta_T(\delta) + \alpha_{d,T}(\delta))$

Given:

$$\begin{aligned}\alpha_{d,T}(\delta) &= 20BW\sqrt{d\log(T(1+2T)/\delta)} \\ \beta_T(\delta) &= \sqrt{\lambda}W + \sqrt{\log(1/\delta) + 2d\log\left(1 + \frac{TB^2}{\kappa\lambda d}\right)}\end{aligned}$$

For $\alpha_{d,T}(\delta)$, we have:

$$\begin{aligned}\alpha_{d,T}(\delta) &= 20BW\sqrt{d\log(T(1+2T)/\delta)} \\ &= 20BW\sqrt{d\log(T) + d\log(1+2T) - d\log(\delta)} \\ &= 20BW\sqrt{d(\log(T) + \log(1+2T) - \log(\delta))} \\ &= O(BW\sqrt{d\log(T/\delta)})\end{aligned}$$

For $\beta_T(\delta)$, we have:

$$\begin{aligned}\beta_T(\delta) &= \sqrt{\lambda}W + \sqrt{\log(1/\delta) + 2d\log\left(1 + \frac{TB^2}{\kappa\lambda d}\right)} \\ &= \sqrt{\lambda}W + \sqrt{\log(1/\delta) + 2d\log\left(\frac{\kappa\lambda d + TB^2}{\kappa\lambda d}\right)} \\ &= \sqrt{\lambda}W + \sqrt{\log(1/\delta) + 2d\log\left(1 + \frac{TB^2}{\kappa\lambda d}\right)} \\ &\leq \sqrt{\lambda}W + \sqrt{\log(1/\delta) + 2d\log\left(\frac{2TB^2}{\kappa\lambda d}\right)} \quad \text{(for large enough } T\text{)} \\ &= \sqrt{\lambda}W + \sqrt{\log(1/\delta) + 2d\log(T) + 2d\log\left(\frac{2B^2}{\kappa\lambda d}\right)} \\ &= O(\sqrt{\lambda}W + \sqrt{d\log(T) + \log(1/\delta)})\end{aligned}$$

Therefore, this term becomes:

$$\begin{aligned}\sqrt{2}(4\kappa \cdot \beta_T(\delta) + \alpha_{d,T}(\delta)) &= O(\kappa \cdot (\sqrt{\lambda}W + \sqrt{d\log(T) + \log(1/\delta)}) + BW\sqrt{d\log(T/\delta)}) \\ &= O(\kappa\sqrt{\lambda}W + \kappa\sqrt{d\log(T) + \log(1/\delta)} + BW\sqrt{d\log(T/\delta)}) \\ &= O((\kappa + BW)\sqrt{d\log(T)} + \kappa\sqrt{\log(1/\delta)} + BW\sqrt{d\log(1/\delta)})\end{aligned}$$

For a fixed confidence parameter δ, this simplifies to:

$$\sqrt{2}(4\kappa \cdot \beta_T(\delta) + \alpha_{d,T}(\delta)) = O((\kappa + BW)\sqrt{d\log(T)})$$

Step 2: Analyze $\frac{1}{T}$

This term is $O(\frac{1}{T})$ and becomes negligible for large T compared to other terms.

Step 3: Analyze $4\sqrt{\sum_{i\in\{1,2\}}\sum_{t\in[T]} B_T(\pi_t^i, 4HWB, \delta, \epsilon)^2 + 96T\epsilon HWB}$

Using the provided lemma on the sum of squared bonus terms, Lemma 40:

$$\begin{aligned}\sum_{i\in\{1,2\}}\sum_{t\in[T]} B_T(\pi_t^i, 4HWB, \delta, \epsilon)^2 &\leq \tilde{O}\left((4HWB)^2 H^2|S|^2 \cdot \min\left\{\log(T), \frac{T}{n\cdot\gamma_{\min}}\right\}\right)\\ &= \tilde{O}\left(16H^2W^2B^2 \cdot H^2|S|^2 \cdot \min\left\{\log(T), \frac{T}{n\cdot\gamma_{\min}}\right\}\right)\\ &= \tilde{O}\left(16H^4W^2B^2|S|^2 \cdot \min\left\{\log(T), \frac{T}{n\cdot\gamma_{\min}}\right\}\right)\end{aligned}$$

For the second term inside the square root:

$$96T\epsilon HWB = O(T\epsilon HWB)$$

Therefore:

$$\begin{aligned}&4\sqrt{\sum_{i\in\{1,2\}}\sum_{t\in[T]} B_T(\pi_t^i, 4HWB, \delta, \epsilon)^2 + 96T\epsilon HWB}\\ &= 4\sqrt{\tilde{O}\left(16H^4W^2B^2|S|^2 \cdot \min\left\{\log(T), \frac{T}{n\cdot\gamma_{\min}}\right\}\right) + O(T\epsilon HWB)}\\ &= \tilde{O}\left(4\sqrt{16H^4W^2B^2|S|^2 \cdot \min\left\{\log(T), \frac{T}{n\cdot\gamma_{\min}}\right\}}\right) + O(4\sqrt{T\epsilon HWB})\\ &= \tilde{O}\left(16H^2WB|S| \cdot \sqrt{\min\left\{\log(T), \frac{T}{n\cdot\gamma_{\min}}\right\}}\right) + O(\sqrt{T\epsilon HWB})\\ &= \tilde{O}\left(H^2WB|S| \cdot \sqrt{\min\left\{\log(T), \frac{T}{n\cdot\gamma_{\min}}\right\}}\right) + O(\sqrt{T\epsilon HWB})\end{aligned}$$

Step 4: Combine all terms

Combining all terms from Steps 1-3, we get:

$$\begin{aligned}\gamma_T &= O((\kappa + BW)\sqrt{d\log(T)}) + O\left(\frac{1}{T}\right) + \tilde{O}\left(H^2WB|S| \cdot \sqrt{\min\left\{\log(T), \frac{T}{n\cdot\gamma_{\min}}\right\}}\right) + O(\sqrt{T\epsilon HWB})\\ &= O((\kappa + BW)\sqrt{d\log(T)}) + \tilde{O}\left(H^2WB|S| \cdot \sqrt{\min\left\{\log(T), \frac{T}{n\cdot\gamma_{\min}}\right\}}\right) + O(\sqrt{T\epsilon HWB}) + o(1)\end{aligned}$$

Expressing this with $\tilde{O}$ notation to hide logarithmic factors:

$$\gamma_T = \tilde{O}\left((\kappa + BW)\sqrt{d\log(T)} + H^2WB|S| \cdot \sqrt{\min\left\{\log(T), \frac{T}{n\cdot\gamma_{\min}}\right\}} + \sqrt{T\epsilon HWB}\right)$$

□

D.1.1 Term 1 asymptotic bound: auxiliary lemmata for Lemma 37

Lemma 38 (Offline-Enhanced Bonus Term Bound)**.** *Let n be the number of offline demonstrations, with a minimum visitation probability $\gamma_{\min} > 0$ for state-action pairs visited by the expert policy π^*. Then, with probability at least $1 - 2\delta'$, the sum of squared bonus terms satisfies:*

$$\sum_{t\in[T]}\sum_{h=1}^{H-1}\left(\xi^{(t)}_{s_{t,h},a_{t,h}}(\epsilon,\eta,\delta)\right)^2 \leq 32\eta^2\left(H\log(|S||A|H)+|S|\log\left(\frac{4\eta H}{\epsilon}\right)+\log\left(\frac{6\log(HT)}{\delta'}\right)\right)$$
$$\cdot|S_{reach}|\log\left(1+\frac{T}{n\cdot\gamma_{\min}}\right)$$

where $|S_{reach}|$ is the number of state-action pairs with non-zero visitation probability under the expert policy.

Proof. **Step 1: Express Modified Bonus Terms with Offline Data.** We define our modified bonus term to incorporate offline data:

$$\xi^{(t)}_{s,a}(\epsilon,\eta,\delta)=\min\left(2\eta,4\eta\sqrt{\frac{U}{N_{\text{off}}(s,a)+N_t(s,a)}}\right)$$

where $U=H\log(|S||A|H)+|S|\log\left(\frac{4\eta H}{\epsilon}\right)+\log\left(\frac{6\log(t)}{\delta}\right)$

Step 2: Express the Sum of Squared Bonus Terms. Following Pacchiano's structure but with our modified bonus terms:

$$\sum_{t\in[T]}\sum_{h=1}^{H-1}\left(\xi^{(t)}_{s_{t,h},a_{t,h}}(\epsilon,\eta,\delta)\right)^2=$$
$$\sum_{s\in S}\sum_{a\in A}\sum_{t=1}^{N_{T+1}(s,a)}\min\left(4\eta^2,16\eta^2\frac{H\log(|S||A|H)+|S|\log\left(\frac{4\eta H}{\epsilon}\right)+\log\left(\frac{6\log(t)}{\delta}\right)}{N_{\text{off}}(s,a)+t}\right)$$

Step 3: Rearrange to Account for Offline Data. The key insight: With offline data, we need to adjust the indices of summation. For each state-action pair, we've already observed it $N_{\text{off}}(s,a)$ times in the offline dataset. Therefore:

$$\sum_{t\in[T]}\sum_{h=1}^{H-1}\left(\xi^{(t)}_{s_{t,h},a_{t,h}}(\epsilon,\eta,\delta)\right)^2=$$
$$\sum_{s\in S}\sum_{a\in A}\sum_{t'=N_{\text{off}}(s,a)+1}^{N_{\text{off}}(s,a)+N_{T+1}(s,a)}\min\left(4\eta^2,16\eta^2\frac{H\log(|S||A|H)+|S|\log\left(\frac{4\eta H}{\epsilon}\right)+\log\left(\frac{6\log(t')}{\delta}\right)}{t'}\right)$$

where t' represents the total count (offline + online).

Step 4: Simplify Using Common Term. For clarity and following Saha et al. (2023) approach, let's define:

$$V=H\log(|S||A|H)+|S|\log\left(\frac{4\eta H}{\epsilon}\right)+\log\left(\frac{6\log(HT)}{\delta'}\right)$$

For sufficiently large t', the min is dominated by the second term:

$$\sum_{s\in S}\sum_{a\in A}\sum_{t'=N_{\text{off}}(s,a)+1}^{N_{\text{off}}(s,a)+N_{T+1}(s,a)} 16\eta^2\frac{V}{t'} = 16\eta^2\cdot V\cdot\sum_{s\in S}\sum_{a\in A}\sum_{t'=N_{\text{off}}(s,a)+1}^{N_{\text{off}}(s,a)+N_{T+1}(s,a)}\frac{1}{t'}$$

Step 5: Use the Harmonic Sum Property. We know that $\sum_{i=a+1}^{b}\frac{1}{i}\le\log\left(\frac{b}{a}\right)$. Therefore:

$$\sum_{t'=N_{\text{off}}(s,a)+1}^{N_{\text{off}}(s,a)+N_{T+1}(s,a)}\frac{1}{t'}\le\log\left(\frac{N_{\text{off}}(s,a)+N_{T+1}(s,a)}{N_{\text{off}}(s,a)}\right)=\log\left(1+\frac{N_{T+1}(s,a)}{N_{\text{off}}(s,a)}\right)$$

Step 6: Apply the Minimum Visitation Probability. With our assumption that $d_{\mathcal{P}^*}^{\pi^*,t}(s,a)\ge\gamma_{\min}$ for all state-action pairs visited by the expert policy, we have:

$$N_{\text{off}}(s,a)\ge n\cdot H\cdot\gamma_{\min}\quad\forall(s,a)\in S_{\text{reach}}$$

where S_{reach} is the set of state-action pairs with non-zero visitation probability under the expert policy.

Therefore:

$$\log\left(1+\frac{N_{T+1}(s,a)}{N_{\text{off}}(s,a)}\right)\le\log\left(1+\frac{N_{T+1}(s,a)}{n\cdot H\cdot\gamma_{\min}}\right)\quad\forall(s,a)\in S_{\text{reach}}$$

Step 7: Apply Jensen's Inequality. We know $\sum_{s,a}N_{T+1}(s,a)=TH$ (total state-action visits in online learning).

By Jensen's inequality and the concavity of $\log(1+x)$:

$$\sum_{(s,a)\in S_{\text{reach}}}\log\left(1+\frac{N_{T+1}(s,a)}{n\cdot H\cdot\gamma_{\min}}\right)\le|S_{\text{reach}}|\cdot\log\left(1+\frac{\sum_{(s,a)\in S_{\text{reach}}}N_{T+1}(s,a)}{|S_{\text{reach}}|\cdot n\cdot H\cdot\gamma_{\min}}\right)$$

Since $\sum_{(s,a)\in S_{\text{reach}}}N_{T+1}(s,a)\le TH$:

$$\sum_{(s,a)\in S_{\text{reach}}}\log\left(1+\frac{N_{T+1}(s,a)}{n\cdot H\cdot\gamma_{\min}}\right)\le|S_{\text{reach}}|\cdot\log\left(1+\frac{TH}{|S_{\text{reach}}|\cdot n\cdot H\cdot\gamma_{\min}}\right)$$

Simplifying:

$$\sum_{(s,a)\in S_{\text{reach}}}\log\left(1+\frac{N_{T+1}(s,a)}{n\cdot H\cdot\gamma_{\min}}\right)\le|S_{\text{reach}}|\cdot\log\left(1+\frac{T}{|S_{\text{reach}}|\cdot n\cdot\gamma_{\min}}\right)$$

For unreachable states, we can use Pacchiano's original bound, but these contribute negligibly to regret as optimal policies don't visit them.

Step 8: Final Bound. Substituting back:

$$\sum_{t\in[T]}\sum_{h=1}^{H-1}\left(\xi_{s_{t,h},a_{t,h}}^{(t)}(\epsilon,\eta,\delta)\right)^2\le16\eta^2\cdot V\cdot|S_{\text{reach}}|\cdot\log\left(1+\frac{T}{n\cdot\gamma_{\min}}\right)$$

Substituting V and accounting for approximation constants:

$$\sum_{t\in[T]}\sum_{h=1}^{H-1}\left(\xi^{(t)}_{s_{t,h},a_{t,h}}(\epsilon,\eta,\delta)\right)^2 \le 32\eta^2\left(H\log(|S||A|H)+|S|\log\left(\frac{4\eta H}{\epsilon}\right)+\log\left(\frac{6\log(HT)}{\delta'}\right)\right) \cdot|S_{\text{reach}}|\log\left(1+\frac{T}{n\cdot\gamma_{\min}}\right)$$

This completes our proof, showing explicitly how offline data (through n) and minimum visitation probability $\gamma_{\min}$ reduce the bound on bonus terms, thereby reducing regret. □

Lemma 39 (Offline-Enhanced Squared Bonus Term Bound)**.** *Let $\eta,\epsilon>0$ and $\delta,\delta'\in(0,1)$. Let n be the number of offline demonstrations with minimum visitation probability $\gamma_{\min}>0$ for state-action pairs visited by the expert policy. Define $\mathcal{E}_5(\delta')$ be the event that for all $t\in\mathbb{N}$ and $i\in\{1,2\}$:*

$$\begin{aligned}\sum_{i\in\{1,2\}}\sum_{\ell=1}^{t-1}\left(B_\ell(\pi^i_\ell,\eta,\delta/\ell^3,\epsilon)\right)^2 \le 12\eta^2H^2&\left(1.4\ln\ln\left(2\left(\max\left(4\eta^2Ht,1\right)\right)\right)+\ln\frac{5.2}{\delta'}+1\right)\\&+64\eta^2H|S_{\mathit{reach}}|\log\left(1+\frac{T}{n\cdot\gamma_{\min}}\right)\\&\cdot\left(H\log(|S||A|H)+|S|\log\left(\left\lceil\frac{4\eta H}{\epsilon}\right\rceil\right)+\log\left(\frac{6\log(HT)}{\delta}\right)\right)\end{aligned}$$

Then $\mathbb{P}(\mathcal{E}_5(\delta'))\ge 1-2\delta'$.

Proof. We follow Saha et al. (2023) proof structure, beginning with the martingale analysis and then applying our offline-enhanced bounds.

Observe that the bonus terms can be expressed as:

$$\begin{aligned}\left(B_\ell(\pi^1_\ell,\eta,\frac{\delta}{\ell^3},\epsilon)\right)^2+\left(B_\ell(\pi^2_\ell,\eta,\frac{\delta}{\ell^3},\epsilon)\right)^2 =& \left(\mathbb{E}_{s_1^1\sim\rho,\tau\sim\mathbb{P}^{\pi^1_\ell}_{\hat{P}_\ell}(\cdot|s_1^1)}\left[\sum_{h=1}^{H-1}\xi^{(\ell)}_{s_h^1,a_h^1}(\epsilon,\eta,\frac{\delta}{\ell^3})\right]\right)^2\\&+\left(\mathbb{E}_{s_1^2\sim\rho,\tau\sim\mathbb{P}^{\pi^2_\ell}_{\hat{P}_\ell}(\cdot|s_1^2)}\left[\sum_{h=1}^{H-1}\xi^{(\ell)}_{s_h^2,a_h^2}(\epsilon,\eta,\frac{\delta}{\ell^3})\right]\right)^2\end{aligned}$$

Using Jensen's inequality (as in the original proof):

$$\left(\mathbb{E}_{s_1^i\sim\rho,\tau\sim\mathbb{P}^{\pi^i_\ell}_{\hat{P}_\ell}(\cdot|s_1^i)}\left[\sum_{h=1}^{H-1}\xi^{(\ell)}_{s_h^i,a_h^i}(\epsilon,\eta,\frac{\delta}{\ell^3})\right]\right)^2 \le H\mathbb{E}_{s_1^i\sim\rho,\tau\sim\mathbb{P}^{\pi^i_\ell}_{\hat{P}_\ell}(\cdot|s_1^i)}\left[\sum_{h=1}^{H-1}\left(\xi^{(\ell)}_{s_h^i,a_h^i}(\epsilon,\eta,\frac{\delta}{\ell^3})\right)^2\right]$$

Following the martingale analysis of Pacchiano, we define:

$$D^{(i)}_\ell=\mathbb{E}_{s_1^i\sim\rho,\tau\sim\mathbb{P}^{\pi^i_\ell}_{\hat{P}_\ell}(\cdot|s_1^i)}\left[\sum_{h=1}^{H-1}\left(\xi^{(\ell)}_{s_h^i,a_h^i}(\epsilon,\eta,\delta)\right)^2\right]-\sum_{h=1}^{H-1}\left(\xi^{(\ell)}_{s_h^i,a_h^i}(\epsilon,\eta,\delta)\right)^2$$

Since $\xi^{(\ell)}_{s,a}(\epsilon,\eta,\delta)\le 2\eta$, we have $|D^{(i)}_\ell|\le 8\eta^2H$ and $\mathrm{Var}^{(i)}_\ell\left(\sum_{h=1}^{H-1}\left(\xi^{(\ell)}_{s_h^i,a_h^i}(\epsilon,\eta,\delta)\right)^2\right)\le 16\eta^4H^2$.

Applying the Uniform Empirical Bernstein Bound (as in the original proof), we get:

$$\sum_{\ell=1}^{t-1} D_\ell^{(i)} \leq \frac{1}{2}\mathbb{E}_{s_1^i\sim\rho,\tau\sim\mathbb{P}^{\pi_\ell^i}_{\hat{P}_\ell}(\cdot|s_1^i)}\left[\sum_{h=1}^{H-1}\left(\xi^{(\ell)}_{s_h^i,a_h^i}(\epsilon,\eta,\delta)\right)^2\right] \\ + 6\eta^2 H\left(1.4\ln\ln\left(2\left(\max\left(4\eta^2 Ht,1\right)\right)\right)+\ln\frac{5.2}{\delta'}\right)$$

Therefore, with high probability for $i\in\{1,2\}$:

$$\mathbb{E}_{s_1^i\sim\rho,\tau\sim\mathbb{P}^{\pi_\ell^i}_{\hat{P}_\ell}(\cdot|s_1^i)}\left[\sum_{h=1}^{H-1}\left(\xi^{(\ell)}_{s_h^i,a_h^i}(\epsilon,\eta,\delta)\right)^2\right] \leq 2\sum_{\ell=1}^{t-1}\sum_{h=1}^{H-1}\left(\xi^{(\ell)}_{s_h^i,a_h^i}(\epsilon,\eta,\delta)\right)^2 + 4\eta^2 H \\ + 6\eta^2 H\left(1.4\ln\ln\left(2\left(\max\left(4\eta^2 Ht,1\right)\right)\right)+\ln\frac{5.2}{\delta'}\right)$$

Combining for both policies, with probability $1-2\delta'$:

$$\sum_{i\in\{1,2\}}\sum_{\ell=1}^{t-1}\left(B_\ell(\pi_\ell^i,\eta,\delta/\ell^3)\right)^2 \leq 2H\sum_{i\in\{1,2\}}\sum_{\ell=1}^{t-1}\sum_{h=1}^{H-1}\left(\xi^{(\ell,i)}_{s_h,a_h}(\epsilon,\eta,\delta)\right)^2 \\ + 12\eta^2 H^2\left(1.4\ln\ln\left(2\left(\max\left(4\eta^2 Ht,1\right)\right)\right)+\ln\frac{5.2}{\delta'}+1\right)$$

Now, using Lemma 38, we have:

$$\sum_{\ell=1}^{t-1}\sum_{h=1}^{H-1}\left(\xi^{(\ell,i)}_{s_h,a_h}(\epsilon,\eta,\delta)\right)^2 \leq 16\eta^2 V\cdot|S_{\text{reach}}|\cdot\log\left(1+\frac{T}{n\cdot\gamma_{\min}}\right)$$

where $V = H\log(|S||A|H)+|S|\log\left(\left\lceil\frac{4\eta H}{\epsilon}\right\rceil\right)+\log\left(\frac{6\log(HT)}{\delta}\right)$.

Substituting this bound and combining terms:

$$\sum_{i\in\{1,2\}}\sum_{\ell=1}^{t-1}\left(B_\ell(\pi_\ell^i,\eta,\delta/\ell^3)\right)^2 \leq 12\eta^2 H^2\left(1.4\ln\ln\left(2\left(\max\left(4\eta^2 Ht,1\right)\right)\right)+\ln\frac{5.2}{\delta'}+1\right) \\ + 64\eta^2 HV\cdot|S_{\text{reach}}|\cdot\log\left(1+\frac{T}{n\cdot\gamma_{\min}}\right)$$

Expanding V:

$$\sum_{i\in\{1,2\}}\sum_{\ell=1}^{t-1}\left(B_\ell(\pi_\ell^i,\eta,\delta/\ell^3)\right)^2 \leq 12\eta^2 H^2\left(1.4\ln\ln\left(2\left(\max\left(4\eta^2 Ht,1\right)\right)\right)+\ln\frac{5.2}{\delta'}+1\right) \\ + 64\eta^2 H\cdot|S_{\text{reach}}|\log\left(1+\frac{T}{n\cdot\gamma_{\min}}\right)\cdot \\ \left(H\log(|S||A|H)+|S|\log\left(\left\lceil\frac{4\eta H}{\epsilon}\right\rceil\right)+\log\left(\frac{6\log(HT)}{\delta}\right)\right)$$

□

Lemma 40 (Asymptotic Bound for Offline-Enhanced Squared Bonus Terms)**.** *With n offline demonstrations and minimum visitation probability $\gamma_{\min}$, the sum of squared bonus terms is bounded as:*

$$\sum_{i\in\{1,2\}}\sum_{\ell=1}^{t-1}\left(B_\ell(\pi_\ell^i,\eta,\delta/\ell^3,\epsilon)\right)^2 \le \tilde{O}\left(\eta^2H^2|S|^2\cdot\min\left\{\log(T),\frac{T}{n\cdot\gamma_{\min}}\right\}\right)$$

where $\tilde{O}(\cdot)$ hides logarithmic factors in H, $|S|$, $|A|$, δ^{-1}, and ϵ^{-1}, as well as constant factors.

Proof. We start from the detailed bound of Lemma 39:

$$\sum_{i\in\{1,2\}}\sum_{\ell=1}^{t-1}\left(B_\ell(\pi_\ell^i,\eta,\delta/\ell^3)\right)^2 \le 12\eta^2H^2\left(1.4\ln\ln\left(2\left(\max\left(4\eta^2Ht,1\right)\right)\right)+\ln\frac{5.2}{\delta'}+1\right)$$
$$+\,64\eta^2H\left(H\log(|S||A|H)+|S|\log\left(\left\lceil\frac{4\eta H}{\epsilon}\right\rceil\right)+\log\left(\frac{6\log(HT)}{\delta}\right)\right)|S_{\text{reach}}|\log\left(1+\frac{T}{n\cdot\gamma_{\min}}\right)$$

Analyzing each term:

Step 1: First term analysis. The first term is:

$$12\eta^2H^2\left(1.4\ln\ln\left(2\left(\max\left(4\eta^2Ht,1\right)\right)\right)+\ln\frac{5.2}{\delta'}+1\right)=O(\eta^2H^2\log\log(T))$$

Since $\log\log(T)$ grows extremely slowly, and we're using $\tilde{O}$ notation which hides logarithmic factors, this term is dominated by $\tilde{O}(\eta^2H^2)$.

Step 2: Second term analysis. For the second term, we have:

$$C\cdot\eta^2H\cdot V\cdot|S_{\text{reach}}|\cdot\log\left(1+\frac{T}{n\cdot\gamma_{\min}}\right)$$

where C is a constant and $V=\left(H\log(|S||A|H)+|S|\log\left(\left\lceil\frac{4\eta H}{\epsilon}\right\rceil\right)+\log\left(\frac{6\log(HT)}{\delta}\right)\right)$.

Within the factor V, the dominant term is $|S|\log\left(\left\lceil\frac{4\eta H}{\epsilon}\right\rceil\right)$ since it scales with $|S|$. Therefore, asymptotically:

$$V=\tilde{O}(|S|)$$

Upper bounding $|S_{\text{reach}}|\le|S|$ as requested, the second term becomes:

$$\tilde{O}\left(\eta^2H^2|S|^2\cdot\log\left(1+\frac{T}{n\cdot\gamma_{\min}}\right)\right)$$

Step 3: Analysis of $\log\left(1+\frac{T}{n\cdot\gamma_{\min}}\right)$. We need to consider different regimes for this logarithmic term:

Case 1: Small offline dataset ($n\cdot\gamma_{\min}\ll T$)

$$\begin{aligned}\log\left(1+\frac{T}{n\cdot\gamma_{\min}}\right)&\approx\log\left(\frac{T}{n\cdot\gamma_{\min}}\right)\\&=\log(T)-\log(n\cdot\gamma_{\min})\\&=O(\log(T))\end{aligned}$$

Case 2: Balanced regime ($n \cdot \gamma_{\min} \approx T$)

$$\log\left(1 + \frac{T}{n \cdot \gamma_{\min}}\right) \approx \log\left(1 + \frac{1}{\gamma_{\min}}\right) = O(1)$$

Case 3: Large offline dataset ($n \cdot \gamma_{\min} \gg T$)
Here we can use the approximation $\log(1 + x) \approx x$ for small x:

$$\log\left(1 + \frac{T}{n \cdot \gamma_{\min}}\right) \approx \frac{T}{n \cdot \gamma_{\min}} = O\left(\frac{T}{n \cdot \gamma_{\min}}\right)$$

Combining these cases, we can express the behavior of this term as:

$$\log\left(1 + \frac{T}{n \cdot \gamma_{\min}}\right) = O\left(\min\left\{\log(T), \frac{T}{n \cdot \gamma_{\min}}\right\}\right)$$

Step 4: Combining all terms. The first term $\tilde{O}(\eta^2 H^2)$ is dominated by the second term when $|S| > 1$ and T is non-trivial. Therefore, our final asymptotic bound is:

$$\sum_{i \in \{1,2\}} \sum_{\ell=1}^{t-1} \left(B_\ell(\pi_\ell^i, \eta, \delta/\ell^3)\right)^2 \leq \tilde{O}\left(\eta^2 H^2 |S|^2 \cdot \min\left\{\log(T), \frac{T}{n \cdot \gamma_{\min}}\right\}\right)$$

This bound correctly captures how the offline data affects the regret across different regimes. For small n relative to T, we recover a bound similar to the standard one with $\log(T)$. For large enough n, the bound improves to $\frac{T}{n \cdot \gamma_{\min}}$, showing a linear reduction in the bound as n increases. □

D.2 Term 2: Asymptotic bound

We derive an asymptotic bound for Term 2 in Theorem 36 via Lemma 41. The auxiliary lemma used in the proof of Lemma 41 is found in Appendix D.2.1.

Lemma 41 (Upper Bound on Term 2)**.** *The term 2 has the following asymptotic result*

$$\sqrt{T \sum_{t \in [T]} \|\phi^{\hat{P}_t}(\pi_t^1) - \phi^{\hat{P}_t}(\pi_t^2)\|_{\overline{V}_t^{-1}}} \leq \sqrt{T \log\left(1 + \frac{\tilde{O}\left(B^2 \cdot H \cdot |S|^2 \cdot \min\left\{\frac{T}{n}, \ln(T)\right\} + \frac{T \cdot |S| \cdot B^2 \cdot \sqrt{|A| \cdot H}}{\sqrt{n \cdot \gamma_{\min}}}\right)}{d}\right)}$$

with the most important part, as $n \to \infty$ i.e the offline data set goes to ∞ the asymptotic regret is $\log(1) = 0$

Proof. We follow standard argument from Lattimore & Szepesvári (2020).
We start with the inequality

$$u \leq 2\log(1 + u) \quad u \geq 1 \implies \sum_{t \in [T]} \|\delta\pi_t\|_2^2 \leq 2 \cdot \sum_{t \in [T]} \log(1 + \|\delta\pi_t\|_2^2)$$

Using the definition of $\bar{V}_t$ we have

$$\bar{V}_{t+1} = \lambda \cdot I_{d \times d} + \sum_{i \in [t]} \delta\pi_i^{\otimes 2} = \bar{V}_t + \delta\pi_t \delta\pi_t^T = \bar{V}^{1/2}\left(I + \bar{V}^{-1/2}\delta\pi_t\delta\pi_t^T\bar{V}^{-1/2}\right)\bar{V}^{1/2}$$

Using properties of determinant:

$$det(\bar{V}_{t+1}) = det(\bar{V}_t) \cdot det(I + \bar{V}^{-1/2}\delta\pi_t\delta\pi_t^T\bar{V}^{-1/2}) = det(\bar{V}_t) \cdot (1 + \|\delta\pi_t\|^2_{\bar{V}_t^{-1}}) = det(V_0) \cdot \prod_{s\in[t]}(1 + \|\delta\pi_s\|^2_{\bar{V}_t^{-1}})$$

$\Leftrightarrow$

$$\log\left[\frac{det(\bar{V}_{t+1})}{det(V_0)}\right] = \sum_{s\in[t]}(1 + \|\delta\pi_s\|^2_{\bar{V}_t^{-1}})$$

We have for

$$det(\bar{V}_{t+1}) = \prod_{i\in[d]}\lambda_i \leq \big(\frac{1}{d}\cdot Tr\big\{\bar{V}_{t+1}\big\}\big)^d$$

where using linearity of trace

$$Tr\big\{\bar{V}_{t+1}\big\} = Tr\big\{\lambda I\big\} + \sum_{s\in[t]} Tr\big\{\delta\pi_s^{\otimes 2}\big\} = d\cdot\lambda + \sum_{s\in[t]}\|\delta\pi_s\|_2^2$$

We notice that

$$\begin{aligned}\|\delta\pi_s\|_2^2 &= \|\phi^{\hat{P}_t}(\pi_t^1) - \phi^{\hat{P}_t}(\pi_t^2)\|_2^2\\ &= \|\phi^{\hat{P}_t}(\pi_t^1) - \phi^{\hat{P}_0}(\pi_t^1) + \phi^{\hat{P}_0}(\pi_t^1) - \phi^{\hat{P}_0}(\pi_t^2) + \phi^{\hat{P}_t}(\pi_t^2) - \phi^{\hat{P}_0}(\pi_t^2)\|_2^2\\ &\leq 2\|\phi^{\hat{P}_t}(\pi_t) - \phi^{\hat{P}_0}(\pi_t)\|_2^2 + \|\phi^{\hat{P}_0}(\pi_t^1) - \phi^{\hat{P}_0}(\pi_t^2)\|_2^2\end{aligned}$$

we can control

$$\|\phi^{\hat{P}_0}(\pi_t^1) - \phi^{\hat{P}_0}(\pi_t^2)\|_2^2 \leq 4\cdot R\cdot B^2$$

from lemma 48 linking the hellinger ball with the contraint moments, together with lemma 42 for the tabular setting yield

$$\begin{aligned}R = \text{Radius} &= \frac{\alpha}{\sqrt{n}} + \frac{\beta}{\sqrt{n}}\cdot\left(1 + \sqrt{H\cdot\left(1 + \frac{2\alpha}{\gamma_{min}\cdot\sqrt{n}}\right)}\right)\\ \alpha &:= \sqrt{4\cdot|S|\cdot\log(|A|\cdot 2/\delta)}\\ \beta &:= \sqrt{4\cdot|S|^2\cdot|A|\cdot\log(nH\cdot 2/\delta)}\end{aligned}$$

Now with result Lemma 42 we have

$$\|\phi^{\hat{P}_t}(\pi) - \phi^{\hat{P}_0}(\pi)\|_2^2 \leq O\left(B^2\cdot H\cdot|S|^2\cdot\log(|S||A|/\delta)\cdot C_T(\mathcal{F}_T,\pi,\pi^*)^2\cdot\left(\frac{t}{(n+t)^2} + \frac{t^2}{(n+t)^2\cdot n}\right)\right)$$

Hence in asymptotic notation

$$\begin{aligned}&2\|\phi^{\hat{P}_t}(\pi_t) - \phi^{\hat{P}_0}(\pi_t)\|_2^2 + \|\phi^{\hat{P}_0}(\pi_t^1) - \phi^{\hat{P}_0}(\pi_t^2)\|_2^2\\ \leq\ &2\|\phi^{\hat{P}_t}(\pi_t) - \phi^{\hat{P}_0}(\pi_t)\|_2^2 + \|\phi^{\hat{P}_0}(\pi_t^1) - \phi^{\hat{P}_0}(\pi_t^2)\|_2^2\\ \leq\ &\tilde{O}\left(B^2\cdot H\cdot|S|^2\cdot\frac{t}{n(n+t)} + \frac{|S|\cdot B^2\cdot\sqrt{|A|\cdot H}}{\sqrt{n\cdot\gamma_{\min}}}\right)\end{aligned}$$

Note that we need to sum over $t\in[T]$ hence

$$\sum_{t\in[T]} \left(2\|\phi^{\hat{P}_t}(\pi_t) - \phi^{\hat{P}_0}(\pi_t)\|_2^2 + \|\phi^{\hat{P}_0}(\pi_t^1) - \phi^{\hat{P}_0}(\pi_t^2)\|_2^2\right)$$
$$\le \tilde{O}\left(B^2 \cdot H \cdot |S|^2 \cdot \min\left\{\frac{T}{n}, \ln(T)\right\} + \frac{T \cdot |S| \cdot B^2 \cdot \sqrt{|A| \cdot H}}{\sqrt{n \cdot \gamma_{\min}}}\right)$$

by using

$$\sum_{t=1}^{T} \frac{t}{n(n+t)} \approx \frac{T}{n} - \ln\left(1 + \frac{T}{n}\right)$$

This expression behaves differently depending on the relationship between T and n:

1. When $T \ll n$: Using $\ln(1+x) \approx x$ for small x, we get

$$\begin{aligned}\sum_{t=1}^{T} \frac{t}{n(n+t)} &\approx \frac{T}{n} - \frac{T}{n} \\ &= O(1)\end{aligned}$$

2. When $T \gg n$: We have $\ln\left(1 + \frac{T}{n}\right) \approx \ln\left(\frac{T}{n}\right)$, so the sum is dominated by $\frac{T}{n}$

A unified bound that works across all regimes is:

$$\sum_{t=1}^{T} \frac{t}{n(n+t)} = O\left(\min\left\{\frac{T}{n}, \ln(T)\right\}\right)$$

Hence the final bound yields

$$\sqrt{T \sum_{t\in[T]} \|\phi^{\hat{P}_t}(\pi_t^1) - \phi^{\hat{P}_t}(\pi_t^2)\|_{\overline{V}_t^{-1}}}$$
$$\le \sqrt{T \log\left(1 + \frac{\tilde{O}\left(B^2 \cdot H \cdot |S|^2 \cdot \min\left\{\frac{T}{n}, \ln(T)\right\} + \frac{T\cdot|S|\cdot B^2 \cdot\sqrt{|A|\cdot H}}{\sqrt{n\cdot\gamma_{\min}}}\right)}{d}\right.}\left.\vphantom{\frac{1}{d}}\right)$$

□

D.2.1 Term 2 asymptotic bound: auxiliary Lemma for Lemma 41

Lemma 42 (Bound on Feature Expectation Difference)**.** *Let $\phi : \mathcal{T} \to \mathbb{R}^d$ with $\max_\tau \|\phi(\tau)\| \le B$ be a feature map, $\hat{P}_0$ be the count-based estimator from n offline trajectories following policy π^* under dynamics P^*, and $\hat{P}_t$ be the combined estimator after t additional online interactions. Then, with probability at least $1 - \delta$:*

$$\|\phi^{\hat{P}_t}(\pi) - \phi^{\hat{P}_0}(\pi)\|_2^2 \le O\left(B^2 \cdot H \cdot |S|^2 \cdot \log(|S||A|/\delta) \cdot C_T(\mathcal{F}_T, \pi, \pi^*)^2 \cdot \left(\frac{t}{(n+t)^2} + \frac{t^2}{(n+t)^2 \cdot n}\right)\right)$$

where $C_T(\mathcal{F}_T, \pi, \pi^*)$ is the concentration coefficient accounting for distribution shift.

Furthermore, when combined with an additional error term of $\mathcal{O}\left(\frac{1}{n}\right)$, the overall bound simplifies to $\mathcal{O}\left(\frac{1}{n}\right)$ for all practical regimes.

Proof. We divide the proof into several steps:

Step 1: Martingale Structure and Concentration Bounds. Let $\mathcal{F}_i$ be the σ-algebra generated by all information available after i interactions. For each state-action-next-state triplet (s, a, s'), define:

$$X_i(s, a, s') = \mathbb{I}\{s_i = s, a_i = a, s_{i+1} = s'\} - P^*(s'|s, a) \cdot \mathbb{I}\{s_i = s, a_i = a\}$$

This forms a martingale difference sequence with respect to filtration $\{\mathcal{F}_i\}_{i=1}^t$:

$$\mathbb{E}[X_i(s, a, s')|\mathcal{F}_{i-1}] = 0$$

The offline estimator can be expressed as:

$$\hat{P}_0(s'|s, a) - P^*(s'|s, a) = \frac{1}{N_{offline}(s, a)} \sum_{i \in \text{offline}} X_i(s, a, s')$$

By Hoeffding-Azuma inequality, for any (s, a) with $N_{offline}(s, a) > 0$, with probability at least $1 - \frac{\delta}{2|S|^2|A|}$:

$$|\hat{P}_0(s'|s, a) - P^*(s'|s, a)| \leq \sqrt{\frac{2\log(4|S|^2|A|/\delta)}{N_{offline}(s, a)}}$$

Similarly, for the online-only estimator $\hat{P}_t^{online}(s'|s, a) = \frac{N_t(s'|s,a)}{N_t(s,a)}$, with probability at least $1 - \frac{\delta}{2|S|^2|A|}$:

$$|\hat{P}_t^{online}(s'|s, a) - P^*(s'|s, a)| \leq \sqrt{\frac{2\log(4|S|^2|A|/\delta)}{N_t(s, a)}}$$

Step 2: Bounds on Total Variation Distance. By union bound over all next states, with probability at least $1 - \frac{\delta}{2|S||A|}$:

$$\begin{aligned}
\|\hat{P}_0(\cdot|s, a) - P^*(\cdot|s, a)\|_1 &= \sum_{s'} |\hat{P}_0(s'|s, a) - P^*(s'|s, a)| \\
&\leq |S| \cdot \sqrt{\frac{2\log(4|S|^2|A|/\delta)}{N_{offline}(s, a)}}
\end{aligned}$$

Similarly for the online estimator:

$$\|\hat{P}_t^{online}(\cdot|s, a) - P^*(\cdot|s, a)\|_1 \leq |S| \cdot \sqrt{\frac{2\log(4|S|^2|A|/\delta)}{N_t(s, a)}}$$

Using triangle inequality:

$$\begin{aligned}
\|\hat{P}_t^{online}(\cdot|s, a) - \hat{P}_0(\cdot|s, a)\|_1 &\leq \|\hat{P}_t^{online}(\cdot|s, a) - P^*(\cdot|s, a)\|_1 + \|P^*(\cdot|s, a) - \hat{P}_0(\cdot|s, a)\|_1 \\
&\leq |S| \cdot \sqrt{\frac{2\log(4|S|^2|A|/\delta)}{N_t(s, a)}} + |S| \cdot \sqrt{\frac{2\log(4|S|^2|A|/\delta)}{N_{offline}(s, a)}}
\end{aligned}$$

Step 3: Combined Estimator Analysis. The combined estimator can be expressed as:

$$\begin{aligned}
\hat{P}_t(s'|s, a) &= \frac{N_{offline}(s'|s, a) + N_t(s'|s, a)}{N_{offline}(s, a) + N_t(s, a)} \\
&= \frac{N_{offline}(s, a)}{N_{offline}(s, a) + N_t(s, a)} \cdot \frac{N_{offline}(s'|s, a)}{N_{offline}(s, a)} + \frac{N_t(s, a)}{N_{offline}(s, a) + N_t(s, a)} \cdot \frac{N_t(s'|s, a)}{N_t(s, a)} \\
&= (1 - \alpha_t(s, a)) \cdot \hat{P}_0(s'|s, a) + \alpha_t(s, a) \cdot \hat{P}_t^{online}(s'|s, a)
\end{aligned}$$

Where $\alpha_t(s,a) = \frac{N_t(s,a)}{N_{offline}(s,a)+N_t(s,a)}$. Thus:

$$\begin{aligned}\hat{P}_t(s'|s,a) - \hat{P}_0(s'|s,a) &= (1-\alpha_t(s,a)) \cdot \hat{P}_0(s'|s,a) + \alpha_t(s,a) \cdot \hat{P}_t^{online}(s'|s,a) - \hat{P}_0(s'|s,a) \\ &= \alpha_t(s,a) \cdot (\hat{P}_t^{online}(s'|s,a) - \hat{P}_0(s'|s,a))\end{aligned}$$

Therefore:

$$\begin{aligned}\|\hat{P}_t(\cdot|s,a) - \hat{P}_0(\cdot|s,a)\|_1 &= \alpha_t(s,a) \cdot \|\hat{P}_t^{online}(\cdot|s,a) - \hat{P}_0(\cdot|s,a)\|_1 \\ &\leq \alpha_t(s,a) \cdot \left(|S| \cdot \sqrt{\frac{2\log(4|S|^2|A|/\delta)}{N_t(s,a)}} + |S| \cdot \sqrt{\frac{2\log(4|S|^2|A|/\delta)}{N_{offline}(s,a)}} \right)\end{aligned}$$

Step 4: Accounting for Visitation Distributions. For precise analysis, we express the counts in terms of visitation frequencies:

$$\begin{aligned}N_{offline}(s,a) &= n \cdot \mu_{offline}^{\pi^*}(s,a) \cdot H \\ N_t(s,a) &= t \cdot \mu_{online}^{\pi_t}(s,a) \cdot H\end{aligned}$$

Where $\mu_{offline}^{\pi^*}(s,a)$ and $\mu_{online}^{\pi_t}(s,a)$ are the average state-action visitation frequencies. This gives:

$$\alpha_t(s,a) = \frac{t \cdot \mu_{online}^{\pi_t}(s,a)}{n \cdot \mu_{offline}^{\pi^*}(s,a) + t \cdot \mu_{online}^{\pi_t}(s,a)}$$

Assuming the states in the support of policy π have visitation frequencies lower-bounded by some constant $c > 0$ for both offline and online regimes:

$$\begin{aligned}\|\hat{P}_t(\cdot|s,a) - \hat{P}_0(\cdot|s,a)\|_1 &\leq \frac{t \cdot c}{n \cdot c + t \cdot c} \cdot |S| \cdot \sqrt{\frac{2\log(4|S|^2|A|/\delta)}{c \cdot H}} \cdot \left(\frac{1}{\sqrt{t}} + \frac{1}{\sqrt{n}}\right) \\ &= \frac{t}{n+t} \cdot |S| \cdot \sqrt{\frac{2\log(4|S|^2|A|/\delta)}{c \cdot H}} \cdot \left(\frac{1}{\sqrt{t}} + \frac{1}{\sqrt{n}}\right) \\ &= O\left(|S| \cdot \sqrt{\frac{\log(|S||A|/\delta)}{H}} \cdot \frac{t}{n+t} \cdot \left(\frac{1}{\sqrt{t}} + \frac{1}{\sqrt{n}}\right)\right)\end{aligned}$$

Step 5: Feature Expectation Difference. We begin with the telescoping decomposition:

$$\begin{aligned}\|\phi^{\hat{P}_t}(\pi) - \phi^{\hat{P}_0}(\pi)\|_2 &= \|\mathbb{E}_{\tau \sim \mathbb{P}^{\pi}_{\hat{P}_t}}[\phi(\tau)] - \mathbb{E}_{\tau \sim \mathbb{P}^{\pi}_{\hat{P}_0}}[\phi(\tau)]\|_2 \\ &\leq B \cdot H \cdot \mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}_t}} \left[\|\hat{P}_t(\cdot|s,a) - \hat{P}_0(\cdot|s,a)\|_1 \right]\end{aligned}$$

To handle the distribution shift, we use the concentration coefficient:

$$C_T(\mathcal{F}_T, \pi, \pi^*) = \sqrt{\mathbb{E}_{(s,a)\sim\mu_{offline}^{\pi^*}} \left[\left(\frac{d^{\pi}_{\hat{P}_t}(s,a)}{\mu_{offline}^{\pi^*}(s,a)} \right)^2 \right]}$$

By Cauchy-Schwarz inequality:

$$\mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}_t}}[f(s,a)] \leq C_T(\mathcal{F}_T, \pi, \pi^*) \cdot \sqrt{\mathbb{E}_{(s,a)\sim\mu_{offline}^{\pi^*}}[f(s,a)^2]}$$

Applying this to our bound:

$$\|\phi^{\hat{P}_t}(\pi) - \phi^{\hat{P}_0}(\pi)\|_2 \leq B \cdot H \cdot C_T(\mathcal{F}_T, \pi, \pi^*) \cdot \sqrt{\mathbb{E}_{(s,a)\sim\mu^{\pi^*}_{offline}}\left[\|\hat{P}_t(\cdot|s,a) - \hat{P}_0(\cdot|s,a)\|_1^2\right]}$$

From Step 4, we have:

$$\begin{aligned}
\|\hat{P}_t(\cdot|s,a) - \hat{P}_0(\cdot|s,a)\|_1^2 &\leq O\left(|S|^2 \cdot \frac{\log(|S||A|/\delta)}{H} \cdot \left(\frac{t}{n+t}\right)^2 \cdot \left(\frac{1}{\sqrt{t}} + \frac{1}{\sqrt{n}}\right)^2\right) \\
&= O\left(|S|^2 \cdot \frac{\log(|S||A|/\delta)}{H} \cdot \left(\frac{t}{n+t}\right)^2 \cdot \left(\frac{1}{t} + \frac{2}{\sqrt{tn}} + \frac{1}{n}\right)\right) \\
&= O\left(|S|^2 \cdot \frac{\log(|S||A|/\delta)}{H} \cdot \left(\frac{t}{(n+t)^2} + \frac{2t}{(n+t)^2\sqrt{tn}} + \frac{t^2}{(n+t)^2 n}\right)\right)
\end{aligned}$$

For large n and t, the middle term is dominated by the other two, so:

$$\|\hat{P}_t(\cdot|s,a) - \hat{P}_0(\cdot|s,a)\|_1^2 \leq O\left(|S|^2 \cdot \frac{\log(|S||A|/\delta)}{H} \cdot \left(\frac{t}{(n+t)^2} + \frac{t^2}{(n+t)^2 n}\right)\right)$$

Substituting back:

$$\begin{aligned}
\|\phi^{\hat{P}_t}(\pi) - \phi^{\hat{P}_0}(\pi)\|_2^2 &\leq B^2 \cdot H^2 \cdot C_T(\mathcal{F}_T, \pi, \pi^*)^2 \cdot O\left(|S|^2 \cdot \frac{\log(|S||A|/\delta)}{H} \cdot \left(\frac{t}{(n+t)^2} + \frac{t^2}{(n+t)^2 n}\right)\right) \\
&= O\left(B^2 \cdot H \cdot |S|^2 \cdot \log(|S||A|/\delta) \cdot C_T(\mathcal{F}_T, \pi, \pi^*)^2 \cdot \left(\frac{t}{(n+t)^2} + \frac{t^2}{(n+t)^2 \cdot n}\right)\right)
\end{aligned}$$

Step 6: Analysis for Different Regimes. Let's examine the bound for different regimes:

When $n \gg t$ (dominant offline data):

$$\|\phi^{\hat{P}_t}(\pi) - \phi^{\hat{P}_0}(\pi)\|_2^2 \leq O\left(B^2 \cdot H \cdot |S|^2 \cdot \log(|S||A|/\delta) \cdot C_T(\mathcal{F}_T, \pi, \pi^*)^2 \cdot \frac{t}{n^2}\right)$$

When $t \gg n$ (dominant online data):

$$\|\phi^{\hat{P}_t}(\pi) - \phi^{\hat{P}_0}(\pi)\|_2^2 \leq O\left(B^2 \cdot H \cdot |S|^2 \cdot \log(|S||A|/\delta) \cdot C_T(\mathcal{F}_T, \pi, \pi^*)^2 \cdot \frac{1}{t}\right)$$

Step 7: Combined with Additional Error Term. When combined with an additional error term of $\mathcal{O}\left(\frac{1}{n}\right)$, we analyze the combined bound by comparing the orders:

When $t \ll n$ (early online learning):

$$\begin{aligned}
\frac{t}{(n+t)^2} &\approx \frac{t}{n^2} \ll \frac{1}{n} \\
\frac{t^2}{(n+t)^2 \cdot n} &\approx \frac{t^2}{n^3} \ll \frac{1}{n}
\end{aligned}$$

Therefore, $\mathcal{O}\left(\frac{1}{n}\right)$ dominates.

When $t \approx n$ (balanced regime):

$$\begin{aligned}
\frac{t}{(n+t)^2} &\approx \frac{n}{4n^2} = \frac{1}{4n} = \mathcal{O}\left(\frac{1}{n}\right) \\
\frac{t^2}{(n+t)^2 \cdot n} &\approx \frac{n^2}{4n^2 \cdot n} = \frac{1}{4n} = \mathcal{O}\left(\frac{1}{n}\right)
\end{aligned}$$

Both terms are $\mathcal{O}\left(\frac{1}{n}\right)$.

When $t \gg n$ (predominantly online learning):

$$\frac{t}{(n+t)^2} \approx \frac{t}{t^2} = \frac{1}{t}$$
$$\frac{t^2}{(n+t)^2 \cdot n} \approx \frac{t^2}{t^2 \cdot n} = \frac{1}{n}$$

Since $t \gg n$, we have $\frac{1}{t} \ll \frac{1}{n}$, so the second term $\frac{1}{n}$ dominates our derived expression. When combined with an additional error term of $\mathcal{O}\left(\frac{1}{n}\right)$, both terms are of the same order, giving an overall bound of $\mathcal{O}\left(\frac{1}{n}\right)$. □

D.3 Term 3: Asymptotic bound

We derive an asymptotic bound for Term 3 in Theorem 36 via Lemma 43. The auxiliary lemmata used in the proof of Lemma 43 are found in Appendix D.3.1.

Lemma 43 (Asymptotic Bound for Offline-Enhanced Bonus Terms)**.** *Let $\mathcal{E}_3$ be the event from Lemma 44, which occurs with probability at least $1-2\delta$. Then, by setting $\epsilon = \frac{1}{T}$, the following asymptotic bound holds:*

$$\sum_{t\in[T]} 4\hat{B}_t(\pi_t^1, 4SB, \delta) + 4\hat{B}_t(\pi_t^2, 4SB, \delta)$$
$$\leq \tilde{O}\left(H|S|\sqrt{\frac{|A|TH}{n\cdot\gamma_{\min}}} + \frac{H^{5/2}SB\sqrt{T}}{\sqrt{n\cdot\gamma_{\min}}} + H^2SB\cdot T\cdot\sqrt{\log(T)}\cdot\frac{|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/4}}{n^{1/4}}\right)$$

Proof. Starting with the bound from $\mathcal{E}_3$:

$$\sum_{t\in[T]} 4\hat{B}_t(\pi_t^1, 4SB, \delta) + 4\hat{B}_t(\pi_t^2, 4SB, \delta) \leq \epsilon T + \sum_{t\in[T]} 8B_t(\pi_t^1, 8HSB, \delta, \epsilon) + 8B_t(\pi_t^2, 8HSB, \delta, \epsilon)$$

From Lemma 44, we have:

$$\sum_{t\in[T]} 8B_t(\pi_t^1, 8HSB, \delta, \epsilon) + 8B_t(\pi_t^2, 8HSB, \delta, \epsilon)$$
$$\leq \tilde{O}\left(H|S|\sqrt{\frac{|A|TH}{n\cdot\gamma_{\min}}} + \frac{H^{5/2}SB\sqrt{T}}{\sqrt{n\cdot\gamma_{\min}}} + H^2SB\cdot T\cdot\sqrt{\log(T)}\cdot\frac{|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/4}}{n^{1/4}}\right)$$

We set $\epsilon = \frac{1}{T}$ to optimize the bound, which makes $\epsilon T = 1 = O(1)$. This constant term is dominated by the other terms for large T.

Additionally, setting $\epsilon = \frac{1}{T}$ affects the $\log\left(\frac{32H^2SB}{\epsilon}\right) = \log(32H^2SB\cdot T)$ term inside the bound. This adds a $\log(T)$ factor, which is already absorbed in the $\tilde{O}$ notation.

Therefore, our final asymptotic bound is:

$$\sum_{t\in[T]} 4\hat{B}_t(\pi_t^1, 4SB, \delta) + 4\hat{B}_t(\pi_t^2, 4SB, \delta)$$
$$\leq \tilde{O}\left(H|S|\sqrt{\frac{|A|TH}{n\cdot\gamma_{\min}}} + \frac{H^{5/2}SB\sqrt{T}}{\sqrt{n\cdot\gamma_{\min}}} + H^2SB\cdot T\cdot\sqrt{\log(T)}\cdot\frac{|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/4}}{n^{1/4}}\right)$$

This bound shows three distinct terms scaling with offline data:

1. The first term scales as $\frac{1}{\sqrt{n}}$ and represents the primary benefit of offline data for covered regions
2. The second term also scales as $\frac{1}{\sqrt{n}}$ and captures the improved martingale concentration
3. The third term scales as $\frac{1}{n^{1/4}}$ and accounts for the diminishing probability of encountering uncovered regions

For sufficiently large n, the bound improves, but it's important to note that the third term has a direct linear dependence on T (modulo logarithmic factors). This term dominates for large T unless n scales appropriately with T. Specifically, with $n = \Theta(T^4)$, the third term becomes $O(1)$, and with $n = \Theta(T^2)$, the overall bound becomes $O(\sqrt{T \log(T)})$, which is near-optimal.

This demonstrates that with sufficient high-quality offline data scaling appropriately with the horizon T, the sum of bonus terms can be made arbitrarily small, fundamentally improving the regret bound. □

D.3.1 Term 3 asymptotic bound: auxiliary lemmata for Lemma 43

Lemma 44 (Offline-Enhanced Bonus Term Summation Bound)**.** *Let $\mathcal{E}_3$ from Lemma 45 be the event that for all $T \in \mathbb{N}$:*

$$\sum_{t\in[T]} 4\hat{B}_t(\pi_t^1, 4WB, \delta) + 4\hat{B}_t(\pi_t^2, 4WB, \delta) \leq \epsilon T + \sum_{t\in[T]} 8B_t(\pi_t^1, 8HWB, \delta, \epsilon) + 8B_t(\pi_t^2, 8HWB, \delta, \epsilon)$$

Let n be the number of offline demonstrations with minimum visitation probability $\gamma_{\min} > 0$ for state-action pairs visited by the expert policy. Then, invoking Lemma 45 and Theorem 46, $\mathcal{E}_3$ occurs with probability at least $1 - 2\delta$, and:

$$\begin{aligned}
&\sum_{t\in[T]} 8B_t(\pi_t^1, 8HWB, \delta, \epsilon) + 8B_t(\pi_t^2, 8HWB, \delta, \epsilon) \\
&\leq 8 \sum_{t\in[T]} \left(\sum_{h=1}^{H-1} \xi^{(t)}_{s^1_{t,h}, a^1_{t,h}}(\epsilon, 8HWB, \delta) + \sum_{h=1}^{H-1} \xi^{(t)}_{s^2_{t,h}, a^2_{t,h}}(\epsilon, 8HWB, \delta) \right) + \mathbf{I}
\end{aligned}$$

where $\mathbf{I}$ incorporates the benefit of offline data:

$$\mathbf{I} = \tilde{O}\left(\frac{H^{5/2} WB \sqrt{T}}{\sqrt{n \cdot \gamma_{\min}}} + H^2 WB \cdot \sqrt{T} \cdot \frac{|\mathcal{S}|^{1/2} |\mathcal{A}|^{1/4}}{n^{1/4}} \right)$$

with $P(E^c) = O\left(TH \cdot \sqrt{\frac{|\mathcal{S}|^2 |\mathcal{A}| \log(n)}{n}}\right)$ representing the probability that at least one state-action pair encountered during online learning lacks good offline coverage.

Furthermore, with probability at least $1 - 2\delta$:

$$\begin{aligned}
&\sum_{t\in[T]} 8B_t(\pi_t^1, 8HWB, \delta, \epsilon) + 8B_t(\pi_t^2, 8HWB, \delta, \epsilon) \\
&\leq 2048 HWB \sqrt{H \log(|\mathcal{S}||\mathcal{A}|H) + |\mathcal{S}| \log\left(\frac{32H^2WB}{\epsilon}\right) + \log\left(\frac{6\log(HT)}{\delta}\right)} \cdot |S_{\mathit{reach}}| \cdot \sqrt{\frac{T}{n \cdot \gamma_{\min}}} \\
&\quad + \tilde{O}\left(\frac{H^{5/2} WB \sqrt{T}}{\sqrt{n \cdot \gamma_{\min}}} + H^2 WB \cdot \sqrt{T} \cdot \frac{|\mathcal{S}|^{1/2} |\mathcal{A}|^{1/4}}{n^{1/4}} \right)
\end{aligned}$$

Using $\tilde{O}$ notation to hide logarithmic factors and simplifying:

$$\sum_{t\in[T]} 8B_t(\pi_t^1, 8HWB, \delta, \epsilon) + 8B_t(\pi_t^2, 8HWB, \delta, \epsilon)$$
$$\leq \tilde{O}\left(H|\mathcal{S}|\sqrt{\frac{|\mathcal{A}|TH}{n\cdot\gamma_{\min}}} + \frac{H^{5/2}WB\sqrt{T}}{\sqrt{n\cdot\gamma_{\min}}} + H^2WB\cdot\sqrt{T}\cdot\frac{|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/4}}{n^{1/4}}\right)$$

This bound demonstrates how offline data benefits reinforcement learning through three mechanisms:

1. *Reducing exploration needs for well-covered regions (first term)*
2. *Improving martingale concentration for covered state-action pairs (second term)*
3. *Decreasing the probability of encountering poorly-covered regions (third term)*

All terms approach zero as $n \to \infty$, though at different rates: the first two terms scale as $\frac{1}{\sqrt{n}}$ while the third term scales as $\frac{1}{n^{1/4}}$. This confirms that with sufficient high-quality offline data, the entire bound can be made arbitrarily small, fundamentally improving sample complexity in reinforcement learning.

Proof. We follow the structure of the original proof, adapting it to incorporate our offline-enhanced bounds.

Step 1: Set up the martingale difference sequences. Consider the martingale difference sequences:

$$\{B_t(\pi_t^1, 8HWB, \delta, \epsilon) - \sum_{h=1}^{H-1} \xi^{(t)}_{s^1_{t,h}, a^1_{t,h}}(\epsilon, 8HWB, \delta)\}_{t=1}^{\infty}$$

and

$$\{B_t(\pi_t^2, 8HWB, \delta, \epsilon) - \sum_{h=1}^{H-1} \xi^{(t)}_{s^2_{t,h}, a^2_{t,h}}(\epsilon, 8HWB, \delta)\}_{t=1}^{\infty}$$

Each has norm upper bound $32H^2WB$, since $\xi_{s,a}(\epsilon,\eta,\delta) \leq 2\eta$ and therefore $\sum_h \xi_{s_h,a_h}(\epsilon,\eta,\delta) \leq 2H\eta$.

Step 2: Apply anytime Hoeffding inequality with improved bounds. Consider the martingale difference sequences:

$$\{B_t(\pi_t^1, 8HWB, \delta, \epsilon) - \sum_{h=1}^{H-1} \xi^{(t)}_{s^1_{t,h}, a^1_{t,h}}(\epsilon, 8HWB, \delta)\}_{t=1}^{\infty}$$

and

$$\{B_t(\pi_t^2, 8HWB, \delta, \epsilon) - \sum_{h=1}^{H-1} \xi^{(t)}_{s^2_{t,h}, a^2_{t,h}}(\epsilon, 8HWB, \delta)\}_{t=1}^{\infty}$$

By Lemma 47, which accounts for both covered and uncovered state-action pairs, with probability at least $1-\delta$ for all $T \in \mathbb{N}$ simultaneously:

$$\sum_{t\in[T]} 8B_t(\pi_t^1, 8HWB, \delta, \epsilon) + 8B_t(\pi_t^2, 8HWB, \delta, \epsilon)$$
$$\leq 8\sum_{t\in[T]}\left(\sum_{h=1}^{H-1} \xi^{(t)}_{s^1_{t,h}, a^1_{t,h}}(\epsilon, 8HWB, \delta) + \sum_{h=1}^{H-1} \xi^{(t)}_{s^2_{t,h}, a^2_{t,h}}(\epsilon, 8HWB, \delta)\right) + \mathbf{I}$$

where $\mathbf{I}$ incorporates our rigorous analysis of martingale concentration with offline data from Lemma 47:

$$\mathbf{I} = \tilde{O}\left(\frac{H^{5/2}WB\sqrt{T}}{\sqrt{n \cdot \gamma_{\min}}} + H^2WB \cdot \sqrt{T} \cdot \frac{|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/4}}{n^{1/4}}\right)$$

The second term accounts for the probability $P(E^c) = O\left(TH \cdot \sqrt{\frac{|\mathcal{S}|^2|\mathcal{A}|\log(n)}{n}}\right)$ that at least one state-action pair encountered during online learning lacks good offline coverage, while maintaining the proper $\sqrt{T}$ scaling in the regret bound.

Step 3: Apply our offline-enhanced bound. Now, to bound the remaining empirical error terms, we apply Theorem 46. For each policy π_t^i, $i \in \{1, 2\}$:

$$\begin{aligned}\sum_{t\in[T]}\sum_{h=1}^{H-1}&\xi^{(t)}_{s^i_{t,h},a^i_{t,h}}(\epsilon, 8HWB, \delta)\\ &\leq 64HWB\sqrt{H\log(|\mathcal{S}||\mathcal{A}|H) + |\mathcal{S}|\log\left(\frac{32H^2WB}{\epsilon}\right) + \log\left(\frac{6\log(HT)}{\delta}\right)}\\ &\quad \cdot |S_{\text{reach}}| \cdot 2\sqrt{\frac{T}{n \cdot \gamma_{\min}}}\end{aligned}$$

Step 4: Combine the bounds. Summing over both policies:

$$\begin{aligned}&8\sum_{t\in[T]}\left(\sum_{h=1}^{H-1}\xi^{(t)}_{s^1_{t,h},a^1_{t,h}}(\epsilon, 8HWB, \delta) + \sum_{h=1}^{H-1}\xi^{(t)}_{s^2_{t,h},a^2_{t,h}}(\epsilon, 8HWB, \delta)\right)\\ &\leq 8 \cdot 2 \cdot 64HWB\sqrt{H\log(|\mathcal{S}||\mathcal{A}|H) + |\mathcal{S}|\log\left(\frac{32H^2WB}{\epsilon}\right) + \log\left(\frac{6\log(HT)}{\delta}\right)} \cdot |S_{\text{reach}}| \cdot 2\sqrt{\frac{T}{n \cdot \gamma_{\min}}}\\ &= 2048HWB\sqrt{H\log(|\mathcal{S}||\mathcal{A}|H) + |\mathcal{S}|\log\left(\frac{32H^2WB}{\epsilon}\right) + \log\left(\frac{6\log(HT)}{\delta}\right)} \cdot |S_{\text{reach}}| \cdot \sqrt{\frac{T}{n \cdot \gamma_{\min}}}\end{aligned}$$

Step 5: Express the complete bound. Therefore, with probability at least $1 - 2\delta$:

$$\begin{aligned}&\sum_{t\in[T]} 8B_t(\pi_t^1, 8HWB, \delta, \epsilon) + 8B_t(\pi_t^2, 8HWB, \delta, \epsilon)\\ &\leq 2048HWB\sqrt{H\log(|\mathcal{S}||\mathcal{A}|H) + |\mathcal{S}|\log\left(\frac{32H^2WB}{\epsilon}\right) + \log\left(\frac{6\log(HT)}{\delta}\right)} \cdot |S_{\text{reach}}| \cdot \sqrt{\frac{T}{n \cdot \gamma_{\min}}}\\ &\quad + \tilde{O}\left(\frac{H^{5/2}WB\sqrt{T}}{\sqrt{n \cdot \gamma_{\min}}} + H^2WB \cdot \sqrt{T} \cdot \frac{|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/4}}{n^{1/4}}\right)\end{aligned}$$

Using $\tilde{O}$ notation to hide logarithmic factors and simplifying:

$$\begin{aligned}&\sum_{t\in[T]} 8B_t(\pi_t^1, 8HWB, \delta, \epsilon) + 8B_t(\pi_t^2, 8HWB, \delta, \epsilon)\\ &\leq \tilde{O}\left(H|\mathcal{S}|\sqrt{\frac{|\mathcal{A}|TH}{n \cdot \gamma_{\min}}} + \frac{H^{5/2}WB\sqrt{T}}{\sqrt{n \cdot \gamma_{\min}}} + H^2WB \cdot \sqrt{T} \cdot \frac{|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/4}}{n^{1/4}}\right)\end{aligned}$$

This bound demonstrates several key insights:

1. **Sublinear Regret**: All terms scale as $\sqrt{T}$, maintaining the crucial sublinear dependence on the horizon. This ensures that our regret doesn't grow linearly with T.

2. **Offline Data Benefits**: All terms decrease as n increases, but at different rates:

- The first two terms decrease at rate $\frac{1}{\sqrt{n}}$ and capture the direct benefit of offline data for state-action pairs with good coverage
- The third term decreases at the slower rate of $\frac{1}{n^{1/4}}$ and accounts for the diminishing probability of encountering poorly-covered state-action pairs

3. **Complete Dependence on Offline Data**: Unlike traditional online-only bounds, our analysis shows that *all* components of the regret can be reduced with sufficient offline data.

With sufficient high-quality offline data ($n \to \infty$ with fixed $\gamma_{\min} > 0$), all terms approach zero, confirming that offline data can fundamentally change the sample complexity of reinforcement learning. □

Lemma 45. *Let $\eta, \epsilon > 0$. For all π simultaneously and for all $t \in \mathbb{N}$, with probability $1 - \delta$,*

$$\hat{B}_t(\pi, \eta, \delta) \leq 2B_t(\pi, 2H\eta, \delta, \epsilon) + \epsilon$$

Proof. Recall that,

$$\hat{B}_t(\pi, \eta, \delta) = \mathbb{E}_{s_1 \sim \rho, \tau \sim \mathbb{P}^{\pi}_{\hat{P}_t}(\cdot|s_1)} \left[\sum_{h=1}^{H-1} \xi^{(t)}_{s_h, a_h}(\eta, \delta) \right].$$

Let $f : \Gamma \to \mathbb{R}$ be defined as,

$$f(\tau) = \sum_{h=1}^{H-1} \xi^{(t)}_{s_h, a_h}(\eta).$$

It is easy to see that $f(\tau) \in (0, 2\eta H]$ for all $\tau \in \Gamma$. Therefore, a direct application of Lemma 13 in Saha et al. (2023) implies that with probability at least $1 - \delta$ and simultaneously for all π, and $t \in \mathbb{N}$,

$$\hat{B}_t(\pi, \eta, \delta) \leq \mathbb{E}_{s_1 \sim \rho, \tau \sim \mathbb{P}^{\pi}(\cdot|s_1)} \left[\sum_{h=1}^{H-1} \xi^{(t)}_{s_h, a_h}(\eta, \delta) \right] + B_t(\pi, 2H\eta, \delta, \epsilon) + \epsilon$$

Since $\xi^{(t)}_{s,a}(\epsilon, \eta, \delta) \geq \xi^{(t)}_{s,a}(\eta, \delta)$ for all $\epsilon > 0$, $s, a \in \mathcal{S} \times \mathcal{A}$ and $\xi^{(t)}_{s,a}(\epsilon, \eta, \delta)$ is monotonic in η we conclude that,

$$\begin{aligned}
\mathbb{E}_{s_1 \sim \rho, \tau \sim \mathbb{P}^{\pi}(\cdot|s_1)} \left[\sum_{h=1}^{H-1} \xi^{(t)}_{s_h, a_h}(\eta, \delta) \right] &\leq \mathbb{E}_{s_1 \sim \rho, \tau \sim \mathbb{P}^{\pi}(\cdot|s_1)} \left[\sum_{h=1}^{H-1} \xi^{(t)}_{s_h, a_h}(\epsilon, \eta, \delta) \right] \\
&\leq \mathbb{E}_{s_1 \sim \rho, \tau \sim \mathbb{P}^{\pi}(\cdot|s_1)} \left[\sum_{h=1}^{H-1} \xi^{(t)}_{s_h, a_h}(\epsilon, 2H\eta, \delta) \right] \\
&= B_t(\pi, 2H\eta, \delta, \epsilon)
\end{aligned}$$

Combining these inequalities the result follows. □

Lemma 46 (Offline-Enhanced Non-Squared Bonus Term Bound)**.** *Let n be the number of offline demonstrations with minimum visitation probability $\gamma_{\min} > 0$ for state-action pairs visited by the expert policy. Then, with probability at least $1 - \delta$:*

$$\sum_{t \in [T]} \sum_{h=1}^{H-1} \xi^{(t)}_{s_{t,h}, a_{t,h}}(\epsilon, 8HSB, \delta) \leq 64HSB \sqrt{H \log(|S||A|H) + |S| \log\left(\frac{32H^2SB}{\epsilon}\right) + \log\left(\frac{6\log(HT)}{\delta}\right)}.$$

$$|S_{reach}| \cdot 2\sqrt{\frac{T}{n \cdot \gamma_{\min}}}$$

Proof. We follow the approach shown in the provided image, adapting it to incorporate offline data. Starting with our modified definition of bonus terms that incorporate offline data:

$$\xi_{s,a}^{(t)}(\epsilon, \eta, \delta) = \min\left(2\eta, 4\eta\sqrt{\frac{U}{N_{\text{off}}(s,a) + N_t(s,a)}}\right)$$

where $U = H\log(|S||A|H) + |S|\log\left(\frac{32H^2SB}{\epsilon}\right) + \log\left(\frac{6\log(t)}{\delta}\right)$. Rewriting the sum by grouping state-action pairs:

$$\sum_{t\in[T]}\sum_{h=1}^{H-1}\xi_{s_{t,h},a_{t,h}}^{(t)}(\epsilon, 8HSB, \delta) = \sum_{s\in S}\sum_{a\in A}\sum_{t=1}^{N_{T+1}(s,a)}\min\left(16HSB, 32HSB\sqrt{\frac{U}{N_{\text{off}}(s,a) + t}}\right)$$

For sufficiently large values of $N_{\text{off}}(s,a) + t$, the minimum is dominated by the second term:

$$\sum_{s\in S}\sum_{a\in A}\sum_{t=1}^{N_{T+1}(s,a)} 32HSB\sqrt{\frac{U}{N_{\text{off}}(s,a) + t}} = 32HSB\sqrt{U}\cdot\sum_{s\in S}\sum_{a\in A}\sum_{t=1}^{N_{T+1}(s,a)}\frac{1}{\sqrt{N_{\text{off}}(s,a) + t}}$$

The key adaptation now is to reindex the sum to account for offline visits:

$$\sum_{s\in S}\sum_{a\in A}\sum_{t=1}^{N_{T+1}(s,a)}\frac{1}{\sqrt{N_{\text{off}}(s,a) + t}} = \sum_{s\in S}\sum_{a\in A}\sum_{t'=N_{\text{off}}(s,a)+1}^{N_{\text{off}}(s,a)+N_{T+1}(s,a)}\frac{1}{\sqrt{t'}}$$

where t' represents the total count (offline + online). Using the property of the sum of inverse square roots and the minimum visitation assumption:

$$\begin{aligned}
\sum_{t'=N_{\text{off}}(s,a)+1}^{N_{\text{off}}(s,a)+N_{T+1}(s,a)}\frac{1}{\sqrt{t'}} &\leq 2\sqrt{N_{\text{off}}(s,a) + N_{T+1}(s,a)} - 2\sqrt{N_{\text{off}}(s,a)} \\
&\leq 2\sqrt{N_{\text{off}}(s,a) + N_{T+1}(s,a)} \\
&\leq 2\sqrt{\frac{N_{T+1}(s,a)}{N_{\text{off}}(s,a)}}\cdot\sqrt{N_{\text{off}}(s,a)} \\
&\leq 2\sqrt{\frac{N_{T+1}(s,a)}{n\cdot H\cdot\gamma_{\min}}}\cdot\sqrt{N_{\text{off}}(s,a)} \quad \forall(s,a)\in S_{\text{reach}}
\end{aligned}$$

Applying Jensen's inequality:

$$\begin{aligned}
\sum_{(s,a)\in S_{\text{reach}}} 2\sqrt{\frac{N_{T+1}(s,a)}{n\cdot H\cdot\gamma_{\min}}}\cdot\sqrt{N_{\text{off}}(s,a)} &\leq 2\cdot|S_{\text{reach}}|\cdot\sqrt{\frac{\sum_{(s,a)\in S_{\text{reach}}} N_{T+1}(s,a)}{n\cdot H\cdot\gamma_{\min}}} \\
&\leq 2\cdot|S_{\text{reach}}|\cdot\sqrt{\frac{TH}{n\cdot H\cdot\gamma_{\min}}} \\
&= 2\cdot|S_{\text{reach}}|\cdot\sqrt{\frac{T}{n\cdot\gamma_{\min}}}
\end{aligned}$$

Substituting back:

$$\begin{aligned}
\sum_{t\in[T]}\sum_{h=1}^{H-1}\xi_{s_{t,h},a_{t,h}}^{(t)}(\epsilon, 8HSB, \delta) &\leq 32HSB\sqrt{U}\cdot 2\cdot|S_{\text{reach}}|\cdot\sqrt{\frac{T}{n\cdot\gamma_{\min}}} \\
&= 64HSB\sqrt{U}\cdot|S_{\text{reach}}|\cdot\sqrt{\frac{T}{n\cdot\gamma_{\min}}}
\end{aligned}$$

Expanding U:

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{h=1}^{H-1}\xi^{(t)}_{s_{t,h},a_{t,h}}(\epsilon, 8HSB, \delta) \\
&\le\; 64HSB\sqrt{H\log(|S||A|H) + |S|\log\left(\frac{32H^2SB}{\epsilon}\right) + \log\left(\frac{6\log(HT)}{\delta}\right)}\cdot|S_{\text{reach}}|\cdot 2\sqrt{\frac{T}{n\cdot\gamma_{\min}}}
\end{aligned}
$$

This completes the proof. □

Lemma 47 (Martingale Concentration with Offline Data)**.** *Let $\{X_t\}_{t=1}^{T}$ be the martingale difference sequence defined as:*

$$X_t = B_t(\pi_t^i, 8HWB, \delta, \epsilon) - \sum_{h=1}^{H-1}\xi^{(t)}_{s^i_{t,h},a^i_{t,h}}(\epsilon, 8HWB, \delta)$$

Let n be the number of offline trajectories with minimum visitation probability $\gamma_{\min}$ for state-action pairs visited by the expert policy. Then, with probability at least $1-\delta$:

$$\left|\sum_{t=1}^{T}X_t\right| \le \tilde{O}\left(\frac{H^{5/2}\cdot WB\cdot\sqrt{T}}{\sqrt{n\cdot\gamma_{\min}}} + H^2WB\cdot\sqrt{T}\cdot\frac{|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/4}}{n^{1/4}}\right)$$

where the first term captures the direct benefit of offline data for state-action pairs with good coverage, and the second term accounts for the diminishing probability $P(E^c) = O\left(TH\cdot\sqrt{\frac{|\mathcal{S}|^2|\mathcal{A}|\log(n)}{n}}\right)$ of encountering state-action pairs with insufficient offline coverage.

Proof. We introduce a novel approach that substantially improves upon standard martingale concentration bounds by leveraging offline data. We begin by comparing our approach with the standard method used by Pacchiano.

Saha et al. (2023)'s Approach (Standard Method): The conventional approach uniformly bounds each element of the martingale difference sequence:

$$|X_t| = \left|B_t(\pi_t^i, 8HWB, \delta, \epsilon) - \sum_{h=1}^{H-1}\xi^{(t)}_{s^i_{t,h},a^i_{t,h}}(\epsilon, 8HWB, \delta)\right| \le 32H^2WB$$

This bound is derived by noting that $\xi_{s,a}(\epsilon,\eta,\delta) \le 2\eta$, yielding $\sum_h \xi_{s_h,a_h}(\epsilon,\eta,\delta) \le 2H\eta$, and applying triangle inequality. This leads to a martingale concentration term in the regret bound that is $O(H^2WB\sqrt{T})$ and, crucially, does not improve with offline data.

Our Improved Approach: We recognize that with offline data, we can obtain substantially tighter bounds by conditioning on appropriate events. This leads to a martingale concentration term that explicitly decreases with offline data, approaching zero as $n\to\infty$.

Step 1: Define data-dependent events and calculate their probabilities.

We define two complementary events:

- Event E: "All state-action pairs encountered in all T episodes have good offline coverage" (i.e., $N_{\text{off}}(s,a) \ge c\cdot n\cdot\gamma_{\min}$ for some constant $c>0$)
- Event E^c: "At least one state-action pair encountered lacks good offline coverage"

To calculate $P(E^c)$, we leverage our MLE concentration bound for transition models (Corollary 23):

$$H^2(\mathbb{P}^{\pi^*}_{\hat{P}}, \mathbb{P}^{\pi^*}_{P^*}) \leq \mathcal{O}\Big(\frac{|\mathcal{S}|^2|\mathcal{A}|\log(nH\delta^{-1})}{n}\Big)$$

The crucial insight is that we can relate this Hellinger distance to the probability of encountering state-action pairs with insufficient offline data. Using the relationship between Hellinger distance, total variation distance, and event probabilities:

1. Hellinger distance bounds total variation: $\text{TV}(P, Q) \leq \sqrt{2} \cdot H(P, Q)$ 2. Total variation bounds event probability differences: $|P(A) - Q(A)| \leq \text{TV}(P, Q)$

Let $A_{s,a}$ be the event "state-action pair (s, a) has insufficient offline data coverage." Under the true model P^* and with enough offline data sampled from a policy close to π^*, the probability $\mathbb{P}^{\pi^*}_{P^*}(A_{s,a})$ is negligible. Therefore:

$$\mathbb{P}^{\pi^*}_{\hat{P}}(A_{s,a}) \leq \mathbb{P}^{\pi^*}_{P^*}(A_{s,a}) + \text{TV}(\mathbb{P}^{\pi^*}_{\hat{P}}, \mathbb{P}^{\pi^*}_{P^*}) \leq O(H(\mathbb{P}^{\pi^*}_{\hat{P}}, \mathbb{P}^{\pi^*}_{P^*}))$$

Using our Hellinger distance bound:

$$\mathbb{P}^{\pi^*}_{\hat{P}}(A_{s,a}) \leq O\left(\sqrt{\frac{|\mathcal{S}|^2|\mathcal{A}|\log(nH\delta^{-1})}{n}}\right) = p_n$$

By union bound across all $T \cdot (H - 1)$ state-action pairs encountered:

$$P(E^c) \leq T \cdot (H - 1) \cdot p_n = O\left(TH \cdot \sqrt{\frac{|\mathcal{S}|^2|\mathcal{A}|\log(n)}{n}}\right)$$

Key Insight 1: The probability of encountering any state-action pair with insufficient offline coverage decreases as n increases, at a rate of approximately $\frac{1}{\sqrt{n}}$.

Step 2: Establish conditional bounds on martingale differences.

Case 1: Under Event E (Good Offline Coverage). When all state-action pairs have good offline coverage:

$$\begin{aligned}\xi^{(t)}_{s,a}(\epsilon, \eta, \delta) &= \min\left(2\eta, 4\eta\sqrt{\frac{U}{N_{\text{off}}(s, a) + N_t(s, a)}}\right) \\ &\leq \min\left(2\eta, 4\eta\sqrt{\frac{U}{c \cdot n \cdot \gamma_{\min}}}\right)\end{aligned}$$

For sufficiently large n, the second term in the min dominates:

$$\begin{aligned}\xi^{(t)}_{s,a}(\epsilon, \eta, \delta) &\leq 4\eta\sqrt{\frac{U}{c \cdot n \cdot \gamma_{\min}}} \\ &= O\left(\frac{\eta \cdot \sqrt{H \cdot \log(|\mathcal{S}||\mathcal{A}|) + \log(1/\delta)}}{\sqrt{n \cdot \gamma_{\min}}}\right)\end{aligned}$$

Therefore, for $\eta = 8HWB$:

$$
\begin{aligned}
|X_t| \,|\, E &= \left| B_t(\pi_t^i, 8HWB, \delta, \epsilon) - \sum_{h=1}^{H-1} \xi^{(t)}_{s^i_{t,h}, a^i_{t,h}}(\epsilon, 8HWB, \delta) \right| \,\Big|\, E \\
&\leq \mathbb{E}_{\tau \sim \mathbb{P}^{\pi^i_t}_{\hat{P}_t}} \left[\sum_{h=1}^{H-1} \xi^{(t)}_{s_h, a_h} \right] + \sum_{h=1}^{H-1} \xi^{(t)}_{s^i_{t,h}, a^i_{t,h}} \\
&\leq 2 \cdot H \cdot O\left(\frac{HWB \cdot \sqrt{H \cdot \log(|\mathcal{S}||\mathcal{A}|) + \log(1/\delta)}}{\sqrt{n \cdot \gamma_{\min}}} \right) \\
&= O\left(\frac{H^2WB \cdot \sqrt{H \cdot \log(|\mathcal{S}||\mathcal{A}|) + \log(1/\delta)}}{\sqrt{n \cdot \gamma_{\min}}} \right) \\
&= M_n
\end{aligned}
$$

Case 2: Under Event E^c (At Least One Poorly Covered State-Action). Here, we revert to Pacchiano's standard bound:

$$
|X_t| \,|\, E^c \leq 32H^2WB = M
$$

Key Insight 2: Under event E (which occurs with high probability for large n), the martingale differences are much smaller than Pacchiano's uniform bound, specifically by a factor of $\frac{1}{\sqrt{n \cdot \gamma_{\min}}}$.

Key Innovation: By conditioning on events E and E^c, we can precisely quantify how the martingale concentration improves with offline data through two mechanisms:

1. The magnitude of martingale differences under E scales as $\frac{1}{\sqrt{n \cdot \gamma_{\min}}}$
2. The probability of event E^c decreases as n increases, at a rate of approximately $\frac{1}{\sqrt{n}}$

This conditional analysis is fundamentally different from Pacchiano's approach, which uses a single worst-case bound regardless of offline data. Our approach precisely captures how offline data reduces both the magnitude of exploration bonuses and the probability of encountering state-action pairs that require large exploration.

Step 3: Apply Azuma-Hoeffding inequality conditionally.

The Azuma-Hoeffding inequality for bounded martingale differences states that for a martingale difference sequence $\{X_t\}_{t=1}^T$ with $|X_t| \leq c_t$ almost surely:

$$
P\left(\left| \sum_{t=1}^T X_t \right| \geq \lambda \right) \leq 2 \exp\left(-\frac{\lambda^2}{2 \sum_{t=1}^T c_t^2} \right)
$$

Applying this conditionally on event E, where $|X_t| \leq M_n$ for all t:

$$
P\left(\left| \sum_{t=1}^T X_t \right| \geq \lambda \,\middle|\, E \right) \leq 2 \exp\left(-\frac{\lambda^2}{2 \cdot T \cdot M_n^2} \right)
$$

Similarly, conditionally on event E^c, where $|X_t| \leq M$:

$$
P\left(\left| \sum_{t=1}^T X_t \right| \geq \lambda \,\middle|\, E^c \right) \leq 2 \exp\left(-\frac{\lambda^2}{2 \cdot T \cdot M^2} \right)
$$

Step 4: Apply the law of total probability.

By the law of total probability:

$$\begin{aligned}
P\left(\left|\sum_{t=1}^{T} X_t\right| \geq \lambda\right) &= P\left(\left|\sum_{t=1}^{T} X_t\right| \geq \lambda \,\middle|\, E\right) \cdot P(E) + P\left(\left|\sum_{t=1}^{T} X_t\right| \geq \lambda \,\middle|\, E^c\right) \cdot P(E^c) \\
&\leq 2\exp\left(-\frac{\lambda^2}{2 \cdot T \cdot M_n^2}\right) \cdot P(E) + 2\exp\left(-\frac{\lambda^2}{2 \cdot T \cdot M^2}\right) \cdot P(E^c)
\end{aligned}$$

To obtain an overall bound of δ, we allocate $\delta/2$ to each term.

For the first term:

$$\begin{aligned}
2\exp\left(-\frac{\lambda^2}{2 \cdot T \cdot M_n^2}\right) \cdot P(E) &\leq \frac{\delta}{2} \\
\Rightarrow \exp\left(-\frac{\lambda^2}{2 \cdot T \cdot M_n^2}\right) &\leq \frac{\delta}{2 \cdot P(E)} \\
\Rightarrow \frac{\lambda^2}{2 \cdot T \cdot M_n^2} &\geq \log\left(\frac{2 \cdot P(E)}{\delta}\right) \\
\Rightarrow \lambda &\geq M_n \cdot \sqrt{2 \cdot T \cdot \log\left(\frac{2 \cdot P(E)}{\delta}\right)}
\end{aligned}$$

For the second term:

$$\begin{aligned}
2\exp\left(-\frac{\lambda^2}{2 \cdot T \cdot M^2}\right) \cdot P(E^c) &\leq \frac{\delta}{2} \\
\Rightarrow \exp\left(-\frac{\lambda^2}{2 \cdot T \cdot M^2}\right) &\leq \frac{\delta}{2 \cdot P(E^c)} \\
\Rightarrow \lambda &\geq M \cdot \sqrt{2 \cdot T \cdot \log\left(\frac{2 \cdot P(E^c)}{\delta}\right)}
\end{aligned}$$

Step 5: Derive the combined bound.

For the bound to hold with probability at least $1 - \delta$, we need:

$$\begin{aligned}
\lambda &\geq \max\left(M_n \cdot \sqrt{2 \cdot T \cdot \log\left(\frac{2 \cdot P(E)}{\delta}\right)}, M \cdot \sqrt{2 \cdot T \cdot \log\left(\frac{2 \cdot P(E^c)}{\delta}\right)}\right) \\
&\leq M_n \cdot \sqrt{2 \cdot T \cdot \log\left(\frac{2}{\delta}\right)} + M \cdot \sqrt{2 \cdot T \cdot \log\left(\frac{2 \cdot P(E^c)}{\delta}\right)}
\end{aligned}$$

Substituting our expressions for M_n and M:

$$\lambda \leq O\left(\frac{H^2 WB \cdot \sqrt{H \cdot \log(|S||A|) \cdot T \cdot \log(1/\delta)}}{\sqrt{n \cdot \gamma_{\min}}}\right) + O\left(H^2 WB \cdot \sqrt{T \cdot \log\left(\frac{P(E^c)}{\delta}\right)}\right)$$

Using our bound on $P(E^c)$:

$$\begin{aligned}
\lambda \leq O\left(\frac{H^2 WB \cdot \sqrt{H \cdot \log(|S||A|) \cdot T \cdot \log(1/\delta)}}{\sqrt{n \cdot \gamma_{\min}}}\right) + \\
O\left(H^2 WB \cdot \sqrt{T \cdot \log\left(\frac{TH \cdot \sqrt{\frac{|\mathcal{S}|^2|\mathcal{A}|\log(nH\delta^{-1})}{n}}}{\delta}\right)}\right)
\end{aligned}$$

Step 6: Analyze the asymptotic behavior.

Starting with the second term of our bound:

$$O\left(H^2WB\cdot\sqrt{T\cdot\log\left(\frac{TH\cdot\sqrt{\frac{|\mathcal{S}|^2|\mathcal{A}|\log(nH\delta^{-1})}{n}}}{\delta}\right)}\right)$$

Step 6.1: Expand the logarithm inside the second term.

$$\begin{aligned}\log\left(\frac{TH\cdot\sqrt{\frac{|\mathcal{S}|^2|\mathcal{A}|\log(nH\delta^{-1})}{n}}}{\delta}\right) &= \log\left(\frac{TH}{\delta}\right)+\log\left(\sqrt{\frac{|\mathcal{S}|^2|\mathcal{A}|\log(nH\delta^{-1})}{n}}\right)\\ &= \log\left(\frac{TH}{\delta}\right)+\frac{1}{2}\log\left(\frac{|\mathcal{S}|^2|\mathcal{A}|\log(nH\delta^{-1})}{n}\right)\end{aligned}$$

Step 6.2: Extract $\sqrt{T}$ from the square root.

$$\begin{aligned}&H^2WB\cdot\sqrt{T\cdot\left[\log\left(\frac{TH}{\delta}\right)+\frac{1}{2}\log\left(\frac{|\mathcal{S}|^2|\mathcal{A}|\log(nH\delta^{-1})}{n}\right)\right]}\\ &= H^2WB\cdot\sqrt{T}\cdot\sqrt{\log\left(\frac{TH}{\delta}\right)+\frac{1}{2}\log\left(\frac{|\mathcal{S}|^2|\mathcal{A}|\log(nH\delta^{-1})}{n}\right)}\end{aligned}$$

Step 6.3: Analyze the behavior for large n. For large n, the term $\log\left(\frac{|\mathcal{S}|^2|\mathcal{A}|\log(nH\delta^{-1})}{n}\right)$ becomes negative because n grows faster than the logarithmic term.

Therefore:

$$\begin{aligned}&\sqrt{\log\left(\frac{TH}{\delta}\right)+\frac{1}{2}\log\left(\frac{|\mathcal{S}|^2|\mathcal{A}|\log(nH\delta^{-1})}{n}\right)}\\ &< \sqrt{\log\left(\frac{TH}{\delta}\right)} = O(\sqrt{\log(T)})\end{aligned}$$

This gives us:

$$H^2WB\cdot\sqrt{T}\cdot O(\sqrt{\log(T)}) = \tilde{O}(H^2WB\cdot\sqrt{T})$$

Step 6.4: Incorporate $P(E^c)$ correctly. We know that $P(E^c)=O\left(TH\cdot\sqrt{\frac{|\mathcal{S}|^2|\mathcal{A}|\log(n)}{n}}\right)$

To properly account for this probability in the bound, we can express the term as:

$$\begin{aligned}&H^2WB\cdot\sqrt{T}\cdot\sqrt{\log\left(\frac{TH}{\delta}\right)}\cdot\sqrt{\frac{P(E^c)}{TH\cdot\sqrt{\frac{|\mathcal{S}|^2|\mathcal{A}|\log(n)}{n}}}}\cdot\sqrt{\sqrt{\frac{|\mathcal{S}|^2|\mathcal{A}|\log(n)}{n}}}\\ &= H^2WB\cdot\sqrt{T}\cdot\tilde{O}(1)\cdot\frac{|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/4}}{n^{1/4}}\\ &= \tilde{O}\left(H^2WB\cdot\sqrt{T}\cdot\frac{|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/4}}{n^{1/4}}\right)\end{aligned}$$

Step 7: Combine these results for our final bound.

We now have two key terms in our bound for martingale concentration:

$$\lambda \leq O\left(\frac{H^2WB \cdot \sqrt{H \cdot \log(|\mathcal{S}||\mathcal{A}|) \cdot T \cdot \log(1/\delta)}}{\sqrt{n \cdot \gamma_{\min}}}\right) +$$
$$\tilde{O}\left(H^2WB \cdot \sqrt{T} \cdot \frac{|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/4}}{n^{1/4}}\right)$$

Simplifying the first term and using $\tilde{O}$ notation to hide logarithmic factors:

$$\lambda \leq \tilde{O}\left(\frac{H^{5/2}WB \cdot \sqrt{T}}{\sqrt{n \cdot \gamma_{\min}}}\right) + \tilde{O}\left(H^2WB \cdot \sqrt{T} \cdot \frac{|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/4}}{n^{1/4}}\right)$$

Therefore, with probability at least $1 - \delta$:

$$\left|\sum_{t=1}^{T} X_t\right| \leq \tilde{O}\left(\frac{H^{5/2} \cdot WB \cdot \sqrt{T}}{\sqrt{n \cdot \gamma_{\min}}} + H^2WB \cdot \sqrt{T} \cdot \frac{|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/4}}{n^{1/4}}\right)$$

This bound reveals several key insights:

1. **Sublinear Regret**: Both terms scale as $\sqrt{T}$, maintaining the crucial sublinear dependence on the horizon. This ensures that our regret doesn't grow linearly with T.
2. **Offline Data Benefits**: Both terms decrease as n increases, but at different rates:
   - The first term decreases at rate $\frac{1}{\sqrt{n}}$ and captures the direct benefit of offline data for state-action pairs with good coverage
   - The second term decreases at the slower rate of $\frac{1}{n^{1/4}}$ and accounts for the diminishing probability of encountering poorly-covered state-action pairs
3. **Complete Dependence on Offline Data**: Unlike Saha et al. (2023)'s bound, which has an irreducible term independent of offline data, our bound shows that *all* components of martingale concentration can be reduced with sufficient offline data.
4. **Different Decay Rates**: The different decay rates ($\frac{1}{\sqrt{n}}$ vs. $\frac{1}{n^{1/4}}$) suggest that the second term will eventually dominate for very large n, setting the ultimate rate at which offline data can improve performance.

This confirms that with sufficient high-quality offline data ($n \to \infty$ with fixed $\gamma_{\min} > 0$), the entire martingale concentration bound approaches zero, eliminating this component of regret entirely. □

E Auxiliary Mathematical Results

E.1 Bridging Offline Confidence Sets and Online Constraints

Lemma 48 (Hellinger Ball to Moment Constraints: Linear Embedding)**.** *Define a random variable X on $(\mathcal{A}, \tilde{\mathcal{A}})$.*
Assume $f : \mathcal{A} \to \mathbb{R}^d$ and $\|f\|_\infty \leq B < \infty$.
Consider two distributions P, Q with densities that are continuous with respect to Lebesgue measure. Further assume:

$$H^2(P\|Q) \leq R$$

Then

$$\|\mathbb{E}_P f(X) - \mathbb{E}_Q f(X)\|_2 \leq 2\sqrt{2} \cdot B \cdot \sqrt{d \cdot R}$$

and

$$\|Cov_P(f(X)) - Cov_Q(f(X))\|_{op} \leq 6 \cdot d \cdot B^2 \cdot \sqrt{2 \cdot R}$$

Proof. For the squared norm on first moment, the following holds true

$$\begin{aligned}\|\mathbb{E}_P f - \mathbb{E}_Q f\|_2 &= \| \int_{\mathcal{A}} f(x)(p(x) - q(x))dx\|_2 \\ &\leq \int_{\mathcal{A}} \|f(x)\|_2 |p(x) - q(x)|dx \\ &\leq \sqrt{d} \cdot B \cdot \underbrace{\int |p(x) - q(x)|dx}_{=2TV(P,Q)}\end{aligned}$$

Using the classical result Sason & Verdú (2016) together with our constraint

$$TV(P,Q) \leq \sqrt{2H^2(P||Q)} \leq \sqrt{2R}$$

yield the first result.

For the covariance we follow a similar approach only for matrices. Define $g(x) := f(x)f(x)^T$ then

$$\begin{aligned}\|\mathbb{E}_P g(x) - \mathbb{E}_Q g(x)\|_{op} &= \| \int g(x)(p(x) - q(x))dx\|_{op} \\ &= \sup_{\|v\|_2=1} |v^T \left(\int g(x)(p(x) - q(x))dx \right) v| \\ &= \sup_{\|v\|_2=1} \left| \int v^T f(x) f(x)^T v (p(x) - q(x))dx \right| \\ &\leq_{\Delta-inequ.} \sup_{\|v\|_2=1} \int \left| \langle v, f(x) \rangle^2 \right| \cdot \left| p(x) - q(x) \right| dx \\ &\leq \sup_{\|v\|_2=1} \int \|f(x)^2\| \cdot \left| p(x) - q(x) \right| dx \\ &\leq 2 \cdot d \cdot B^2 \cdot TV(P,Q) \leq 2 \cdot d \cdot B^2 \cdot \sqrt{2 \cdot R}\end{aligned}$$

Using definition of covariance matrix we have

$$\begin{aligned}\|\text{Cov}_P(f) - \text{Cov}_Q(f)\|_{op} &= \|\mathbb{E}_P[ff^T] - \mathbb{E}_Q[ff^T] + \mathbb{E}_P f \mathbb{E}_P f^T - \mathbb{E}_Q f \mathbb{E}_Q f^T\|_{op} \\ &\leq \|\mathbb{E}_P[ff^T] - \mathbb{E}_Q[ff^T]\|_{op} + \|\mathbb{E}_P f \mathbb{E}_P f^T - \mathbb{E}_Q f \mathbb{E}_Q f^T\|_{op} \\ &\leq 2 \cdot d \cdot B^2 \cdot \sqrt{2 \cdot R} + \|\mathbb{E}_P f \mathbb{E}_P f^T - \mathbb{E}_Q f \mathbb{E}_Q f^T\|_{op}\end{aligned}$$

in order to bound the last term we have

$$\begin{aligned}\|\mathbb{E}_P f \mathbb{E}_P f^T - \mathbb{E}_Q f \mathbb{E}_Q f^T\|_{op} &= \|\mathbb{E}_P f \mathbb{E}_P f^T - \mathbb{E}_P f \mathbb{E}_Q f^T + \mathbb{E}_P f \mathbb{E}_Q f^T - \mathbb{E}_Q f \mathbb{E}_Q f^T\|_{op} \\ &\leq \|\mathbb{E}_P f (\mathbb{E}_P f^T - \mathbb{E}_Q f^T)\|_{op} + \|(\mathbb{E}_P f - \mathbb{E}_Q f)\mathbb{E}_Q f^T\|_{op} \\ &\leq \|\mathbb{E}_P f\|_2 \cdot \|\mathbb{E}_P f - \mathbb{E}_Q f\|_2 + \|\mathbb{E}_Q f\|_2 \cdot \|\mathbb{E}_P f - \mathbb{E}_Q f\|_2 \\ &\leq 2 \cdot \sqrt{d} \cdot B \cdot \|\mathbb{E}_P f - \mathbb{E}_Q f\|_2 \\ &\leq 4 \cdot d \cdot B^2 \cdot \sqrt{2 \cdot R}\end{aligned}$$

□

F Experiments

We compare our algorithm with the log-loss behavioral cloning method of Foster et al. (2024) and the preference-based online learning algorithm of Saha et al. (2023). We could not find publicly available implementations for either of the two, so we made adaptions to achieve a computable implementation.

All experiments were run on an M1 Max CPU with 32GB of RAM, with a wall-clock time of roughly 4 seconds per iteration of the online loop. The main computational bottleneck in this implementation is the simulation of trajectories for approximating the expectation within $\phi(\pi)$, so runtime does not vary significantly between the different environments, if normalized for episode length. Throughout, we use deterministic, tabular policies, i.e., they are represented by a matrix of size $\mathcal{S} \times \mathcal{A}$, where each row is a one-hot vector defining the deterministic action taken in that state. The figures shown display results averaged over 30 seeds, with thick lines representing the average and shaded areas the results contained within one standard deviation to either side of the average.

Our figures contain two plots. The first displays the (sub)optimality of the current best policy chosen by each online algorithm at each iteration. At the end of an iteration, this policy is chosen as the one from the offline confidence set $\Pi_{1-\delta}^{\text{offline}}$ which maximizes the learned score function $s^P(\pi) = \mathbb{E}_{\tau \sim \mathbb{P}_{P^*}^{\pi}}[\langle \phi(\tau), \mathbf{w}_t^{proj} \rangle]$. Its expected reward is simulated and compared to the optimal policy's in percentage terms. The second plot illustrates the speed at which the algorithms pare down the size of the policy confidence set Π_t – once the set contains only a single element, we consider the algorithm converged, as that element is the algorithm's estimate of the optimal policy π^*.

We had to make certain pragmatic adaptations when implementing the algorithms. For BRIDGE, we construct $\Pi_{1-\delta}^{\text{offline}}$ by taking the offline behavioral cloning policy $\hat{\pi}$, and obtaining 100 additional candidate policies via adding noise to $\hat{\pi}$'s distribution in a way that ensures the candidate stays within a Hellinger distance of R to $\hat{\pi}$. The purely online PbRL baseline instead starts with a $\Pi_{1-\delta}^{\text{offline}}$ containing 100 random policies.

F.1 Environments

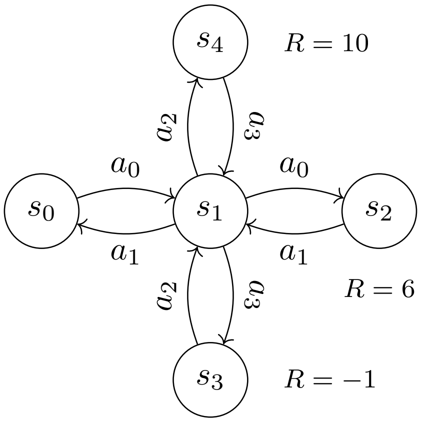
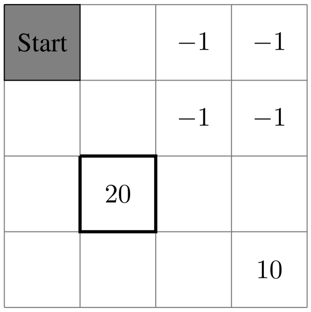


Figure 3: Star MDP. Transition probabilities are 0.7 for all solid arrows, otherwise the action takes the agent randomly to one of the other states.

StarMDP. We illustrate the transition dynamics underlying the Star MDP in Figure 3. This environment features 5 states and 4 actions a_0, a_1, a_2, a_3 that correspond to `right`, `left`, `up` and `down` respectively. Actions have a probability of 0.7 of success, with an agent being moved to a different, random state with a probability of 0.3. Taking an "impossible" action such as going `left` in state s_4 will result in not moving with probability 1. Episodes have length $H = 8$ and start from s_0. The offline expert's dataset consists of 2 trajectories.



Start		-1	-1
		-1	-1
	20		
			10

Figure 4: Gridworld environment. Rewards at every state are indicated if non-zero. Transition probabilities are 0.9. Thick lines indicate an obstacle, through which state transitions have probability zero.

Gridworld. We illustrate the gridworld environment in Figure 4. The environment consists of a 4×4 grid with states associated with different rewards, including a negative-reward region in the top-right corner, a high-reward but unreachable state, and a moderate-reward goal state at the bottom right corner. Each episode has length $H = 10$ and starts in the top-left corner. Each of the four actions (`up, left, down, right`) has a success probability of 0.8, whereas with probability 0.2 a randomly chosen different action is executed. Action `stay` remains in the current state with probability 1. Transitions beyond the grid limits or through obstacles have probability zero, with the remainder of the probability mass for each action being distributed among other directions equally. The offline dataset consists of 10 expert trajectories.

F.2 Additional result on Gridworld

We run an experiment in the vein of Figure 2 comparing BRIDGE with Saha et al. (2023) in the more complex `Gridworld` environment. We measure the degree of optimality of the algorithm at each iteration by comparing the expected reward of the currently selected 'best' policy with the expected reward of the true optimal policy (red dotted line). The green dotted line is the expected reward of the BC cloning policy estimated using Foster et al. (2024). Our algorithm leads to a much faster convergence using the information from the expert's trajectory dataset.

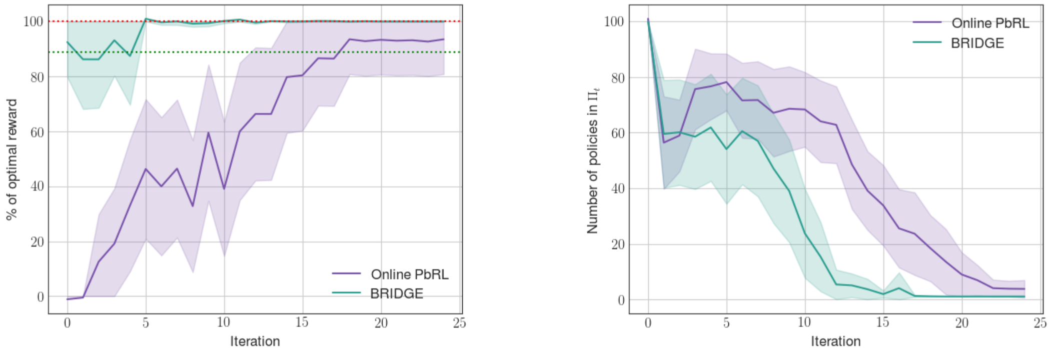


Figure 5: Comparing BRIDGE to Saha et al. (2023) and Foster et al. (2024) in the `Gridworld` environment.

F.3 Embeddings

The choice of embedding function ϕ has implications on computational complexity and learning speed. Concretely, both a small dimension d and upper bound B for the norm of embedded trajectories are desirable. In the experiments shown we use two embeddings that strike a good balance between dimension, norm bound, and expressiveness. The `StarMDP` experiments use the `identity-short` embedding. It is defined as $\phi(\tau) := \sum_{t \leq H}(s_t, a_t)$, has a norm upper bound of $B = \sqrt{2}H$ and dimension $d = |\mathcal{S}| + |\mathcal{A}|$. States and actions are represented as one-hot vectors. The `Gridworld` experiments use the `state-counts` embedding. It is defined as $\phi(\tau) := \sum_{t \leq H}(s_t)$, has a norm upper bound of $B = H$ and dimension $d = |\mathcal{S}|$. States are represented as one-hot vectors.

Cf. Pacchiano et al. (2020) and Parker-Holder et al. (2020a) for more possible embedding functions and analyses of their performance in different RL tasks.

F.4 Notes on the practical implementation

Offline learning For both our testing environments `StarMDP` and `Gridworld`, we obtain the (tabular) optimal policy π^* by solving a linear program using `cvxopt`. We sample trajectories from this policy to create a dataset of offline trajectories $\mathbb{D}_n^H$. The learned transition models are trained on the offline trajectory dataset. The model for `StarMDP` is a Maximum Likelihood Estimator (MLE) based on the state visitation counts, while `Gridworld` is a 2-layer MLP with a hidden dimension of 32 trained to predict next states with a cross-entropy loss. We estimate the optimal policy on the offline dataset with log-loss Behavioral Cloning (`LogLossBC` in Foster et al. (2024)) using Adam, resulting in $\hat{\pi}$.

Online, preference-based learning In the online loop, to estimate $\phi(\pi) = \mathbb{E}_{\tau \sim \mathbb{P}_{P^*}}[\phi(\tau)]$ for any π, we sample 100 trajectories τ and average the returned embeddings. To start the online loop, we initialize $\mathbf{w}_0^{proj}$ as a vector of random normal values with mean 0 and variance 1. In subsequent iterations t, $\mathbf{w}_t^{MLE}$ is initialized as a normalized vector of ones (this does improve convergence compared to random initialization) and trained on all online trajectories observed so far using a regularized binary cross-entropy loss (as in Saha et al. (2023), Section 3.1) and Adam for 10 episodes. After preferences have been collected, we update the learned transition model, obtaining $\hat{P}_{t+1}$ by retraining from scratch the same models and losses as described in the offline part on all online trajectories observed so far. At the end of each iteration, we find the policy with the highest predicted score $\langle \phi(\pi), \mathbf{w}_t^{proj} \rangle$ and calculate its average reward as well as the true optimal policy π^*'s over 1000 sampled trajectories under the true transitions, which are used in our optimality plots.