# ADVERSARIAL REINFORCEMENT LEARNING WITH TRANSITION PREDICTIONS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We study episodic adversarial Markov decision processes with unknown transition dynamics, where the learner is additionally given a potentially misspecified prediction of the transition kernel. We propose Transition-Prediction Follow-the-Regularized-Leader (TP-FTRL), an occupancy-based algorithm that combines empirical transition confidence sets with likelihood-calibrated confidence sets induced by the prediction. We establish simultaneous high-probability coverage of the true transition kernel without assuming prediction accuracy. We further prove a prediction-independent worst-case regret bound of $\tilde{\mathcal{O}}(K|\mathcal{S}|\sqrt{|\mathcal{A}|T})$, matching the state-of-the-art bound by Jin et al. (2020), and showing that arbitrarily inaccurate transition predictions do not degrade the worst-case regret guarantee. When the prediction is close to the empirical transition estimates, the likelihood-calibrated constraints tighten transition uncertainty and yield a sharper regret bound. These results show that TP-FTRL can exploit informative transition predictions while remaining robust to misspecification.

## 1 INTRODUCTION

Markov decision processes (MDPs) (Puterman, 2014) provide a standard mathematical framework for reinforcement learning (Sutton et al., 1998) and sequential decision making in stochastic environments. Adversarial reinforcement learning extends this framework to settings in which components of the environment, such as rewards, losses, or transition dynamics, may vary nonstationarily or adversarially over time (Even-Dar et al., 2009; Zimin & Neu, 2013; Abbasi Yadkori et al., 2013; Dick et al., 2014). In this work, we consider a $K$-step episodic adversarial MDP over $T$ episodes with state space $\mathcal{S}$, action space $\mathcal{A}$, a fixed but unknown transition kernel $P$, and adversarial losses. When the transition kernel is unknown, the learner must adapt to changing losses while learning the underlying dynamics. However, transition information may already be available before interaction through historical observations, simulators, previously learned models, or related environments (Lyu et al., 2025). Recent work has begun to investigate how such transition predictions can improve learning efficiency in stationary discounted MDPs (Lyu et al., 2025; Lu et al., 2026). In practice, their accuracy may be unknown, and the predictions can be misspecified. This raises a natural question:

*Can a learner exploit informative transition predictions while retaining reliable guarantees when they are inaccurate?*

Existing work on adversarial MDPs with unknown transitions primarily learns the dynamics from interaction by constructing empirical transition confidence sets and incorporating them into occupancy-based policy optimization (Rosenberg & Mansour, 2019a; Jin et al., 2020; 2021; 2022; Zimin & Neu, 2013). Under bandit feedback, transition uncertainty is particularly important because it also enters the estimation of unobserved losses through upper occupancy bounds (Jin et al., 2020), which achieves the best regret of $\tilde{\mathcal{O}}(K|\mathcal{S}|\sqrt{|\mathcal{A}|T})$. Likelihood-ratio methods provide a principled framework for constructing sequential confidence sets using predictive information without requiring the prediction itself to be correct (Emmenegger et al., 2023). Existing studies have considered adversarial MDPs with unknown transitions and bandit feedback, as well as learning with transition predictions in stationary discounted MDPs (Lyu et al., 2025; Lu et al., 2026), but these directions have largely been studied separately. To our knowledge, it remains unclear how to incor-



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

porate potentially misspecified transition predictions into episodic adversarial MDPs with unknown transitions and bandit feedback while preserving a prediction-independent regret guarantee.

To address this gap, we study episodic adversarial MDPs with unknown fixed transitions and a potentially misspecified transition prediction. We propose Transition-Prediction Follow-the-Regularized-Leader (TP-FTRL), which combines empirical and likelihood-calibrated confidence sets within occupancy-based FTRL. TP-FTRL preserves a prediction-independent worst-case regret guarantee $\tilde{\mathcal{O}}(K|\mathcal{S}|\sqrt{|\mathcal{A}|T})$ while achieving a sharper bound when the discrepancy between the empirical transition estimate and the transition prediction is small.

### 1.1 Challenges

**Potentially Misspecified Transition Predictions.** A key challenge is that the accuracy of a transition prediction cannot be directly assessed when the true transition kernel is unknown. Recent work characterizes performance using the discrepancy between predicted and true transitions, even when this error is unknown to the learner (Lyu et al., 2025). Our analysis instead compares the fixed transition prediction with the empirical transition distribution estimated from observed transitions. We use KL divergence to quantify this difference, providing a natural likelihood-based measure for transition distributions (Agrawal, 2020; Emmenegger et al., 2023; Kirschner et al., 2025).

**Transition uncertainty under bandit feedback.** Unknown transitions affect both policy optimization and loss estimation through upper occupancy bounds. Existing work addresses this difficulty by constructing empirical transition confidence sets and using upper occupancy bounds to obtain valid bandit loss estimators (Jin et al., 2020; 2021; 2022). In our setting, transition uncertainty is additionally influenced by a potentially misspecified transition prediction through the likelihood-calibrated confidence set. We therefore need to incorporate the prediction into both occupancy-based policy optimization and upper-occupancy loss estimation while retaining high-probability coverage of the true transition kernel under prediction misspecification.

**Exploiting predictions without sacrificing worst-case guarantees.** A desirable learning method should benefit from informative transition predictions while retaining reliable guarantees under prediction misspecification. Prior work shows that imperfect transition predictions can improve learning efficiency in stationary discounted MDPs (Lyu et al., 2025; Lu et al., 2026). We instead study episodic adversarial MDPs with unknown transitions and bandit feedback, and seek to exploit transition predictions to improve learning while preserving a prediction-independent worst-case regret guarantee.

### 1.2 Contributions

Our contributions are summarized as follows.

1. **Method and tractability.** We propose TP-FTRL, which combines empirical and likelihood-calibrated transition confidence sets to safely exploit potentially misspecified transition predictions. We further show that the TP-FTRL update admits an equivalent convex formulation over extended occupancy measures.
2. **Prediction-independent guarantee.** We establish simultaneous high-probability coverage of the true transition kernel without assuming prediction accuracy and prove the prediction-independent regret worst-case bound $\tilde{\mathcal{O}}\Big(K|\mathcal{S}|\sqrt{|\mathcal{A}|T}\Big)$, matching the state-of-the-art bound by Jin et al. (2020).
3. **Discrepancy-dependent guarantee.** We introduce the maximum normalized discrepancy ratio $\kappa$ between the empirical transition estimate and the transition prediction and establish $\tilde{\mathcal{O}}\Big(K^{3/2}\sqrt{(1+\kappa)|\mathcal{S}||\mathcal{A}|T}\Big)$, explicitly quantifying how the regret guarantee varies with such discrepancy.
4. **Unified guarantee.** Combining the prediction-independent and discrepancy-dependent analysis yields $\tilde{\mathcal{O}}\Big(K\sqrt{|\mathcal{S}||\mathcal{A}|T}\min\Big\{\sqrt{K(1+\kappa)},\ \sqrt{|\mathcal{S}|}\Big\}\Big)$.

## 2 Related Work

**Adversarial MDPs with unknown transitions.** A broad line of work studies episodic adversarial MDPs with unknown dynamics under full-information or bandit feedback, typically combining transition estimation or confidence sets with occupancy-based optimization and, under bandit feedback, specialized loss estimators or occupancy envelopes (Rosenberg & Mansour, 2019a;b; Jin et al., 2020; 2021; 2022; Li et al., 2024b;a; Cheng et al., 2026; Chen & Luo, 2021; Lancewicki et al., 2023; Liu et al., 2024b; Wang et al., 2026). Recent extensions also consider structured models such as linear-mixture and low-rank MDPs (Li et al., 2024b; Liu et al., 2024a; Cheng et al., 2026; Zhao et al., 2023; Kong et al., 2023; Dai et al., 2023). Our work builds on adversarial MDPs with unknown transitions and bandit feedback, and further incorporates a fixed transition prediction.

**Learning with prior transition information and predictions.** Another line of work exploits prior information about transition dynamics through predicted transition matrices, multi-step transition predictions, Bayesian priors, or finite sets of candidate transition models (Lyu et al., 2025; Lu et al., 2026; Gopalan & Mannor, 2015; Li & Luo, 2024; Tirinzoni et al., 2020; Nouri, 2009). These works demonstrate the value of additional transition information, but they largely focus on stationary discounted, Bayesian, or robust-MDP settings rather than episodic adversarial MDPs with unknown transitions and bandit feedback. Our setting instead considers a potentially misspecified fixed transition prediction and derives guarantees in terms of its discrepancy from empirical transition estimates.

## 3 Preliminaries

We consider an episodic loop-free adversarial Markov decision process, defined by a tuple $\mathcal{M} = \big(\mathcal{S}, \mathcal{A}, P, K, \{\ell_t\}_{t=1}^T\big)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ is the transition kernel, and $\ell_t : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the loss function in episode $t$. Following the loop-free layered structure, we assume that the state space is partitioned into $K+1$ disjoint layers $\mathcal{S} = \bigcup_{k=0}^K \mathcal{S}_k$, with $\mathcal{S}_i \cap \mathcal{S}_j = \emptyset$ for any $i \neq j$. Transitions are allowed only between consecutive layers; that is, for each $k \in \{0, 1, \dots, K-1\}$ and $P(s'|s,a) = 0$, $\forall s \in \mathcal{S}_k, a \in \mathcal{A}, s' \notin \mathcal{S}_{k+1}$. $\mathcal{S}_0$ and $\mathcal{S}_K$ are singletons.

The learner interacts with the environment for $T$ episodes. In each episode $t$, the environment decides a loss function $\ell_t$, which may depend on the learner's algorithm and all randomness revealed before episode $t$, but not on the learner's randomization within episode $t$. The learner simultaneously starts from state $s_0$ and selects a stochastic policy $\pi_t : \mathcal{S} \times \mathcal{A} \to [0,1]$, where $\pi_t(a|s)$ is the probability of taking action $a$ at a given state $s$ such that $\sum_{a \in \mathcal{A}} \pi_t(a|s) = 1$ for any state $s$. The learner generates a trajectory under this policy, $(s_0, a_0, \dots, s_{K-1}, a_{K-1})$. For each layer $k$, action $a_k$ is drawn from $\pi_t(\cdot|s_k)$ and the next state $s_{k+1}$ is drawn from $P(\cdot|s_k, a_k)$. After each episode, the learner only observes the losses for the visited state-action pairs (*bandit feedback losses*).

Before interaction, the learner is given a prior transition prediction $\tilde{P}$, which has full support on each next-state layer but may otherwise be arbitrarily misspecified. The true transition kernel $P$ remains unknown and is learned from observed trajectories.

For a policy $\pi$ and the transition kernel $P$, we define the *occupancy measure* $q^{P,\pi} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ by $q^{P,\pi}(s,a,s') = \Pr\{s_k = s, a_k = a, s_{k+1} = s' | P, \pi\}$ for $s \in \mathcal{S}_k$, $a \in \mathcal{A}$, and $s' \in \mathcal{S}_{k+1}$. We also denote $q^{P,\pi}(s,a) = \Pr\{s_k = s, a_k = a | P, \pi\} = \sum_{s' \in \mathcal{S}_{k+1}} q^{P,\pi}(s,a,s')$ for $s \in \mathcal{S}_k$ and $a \in \mathcal{A}$. The expected loss following a policy $\pi$ in episode $t$ is $V^{P,\pi}(\ell_t) = \sum_{k=0}^{K-1} \sum_{s \in \mathcal{S}_k} \sum_{a \in \mathcal{A}} q^{P,\pi}(s,a) \ell_t(s,a) = \langle q^{P,\pi}, \ell_t \rangle$. We then define the *regret*:

$$Reg_T(\pi^*) = \sum_{t=1}^T \Big[ V^{P,\pi_t}(\ell_t) - V^{P,\pi^\star}(\ell_t) \Big] = \sum_{t=1}^T \langle q^{P,\pi_t} - q^{P,\pi^*}, \ell_t \rangle,$$

where $\pi^* \in \arg\min_\pi \sum_{t=1}^T V^{P,\pi}(\ell_t)$.

# 4 LEARNING WITH TRANSITION PREDICTIONS

We develop an occupancy-based learning framework, shown in Algorithm 1, that incorporates a possibly misspecified transition prediction through likelihood-calibrated confidence sets. The framework retains empirical transition estimation and adjusts the restrictions induced by the prediction using observed transitions. Our policy update follows the occupancy-based FTRL framework (McMahan, 2017; Cesa-Bianchi & Orabona, 2021; Shalev-Shwartz, 2025), while the loss estimator retains the upper-occupancy implicit-exploration construction.

**Algorithm 1:** TP-FTRL

**Input:** Prior transition prediction $\tilde{P}$, state space $\mathcal{S}$, action space $A$, episodes $T$, confidence level $\delta$, learning rate $\eta > 0$, and implicit-exploration parameter $\gamma > 0$.

**Initialization:**

For all $k = 0, 1, \ldots, K-1$, $(s, a, s') \in \mathcal{S}_k \times \mathcal{A} \times \mathcal{S}_{k+1}$, initialize all transition counts to zero, $\mathcal{C}_1$ to the set of all valid transition functions, and occupancy measure $\hat{q}_1(s, a, s') = \frac{1}{|\mathcal{S}_k||\mathcal{A}||\mathcal{S}_{k+1}|}$. Initialize $\pi_1$ by Equation (3).

**for** $t = 1, \ldots, T$ **do**

- Execute policy $\pi_t$; observe trajectory $\{(s_{t,k}, a_{t,k})\}_{k=0}^{K-1}$ together with the losses at the visited state-action pairs.
- Compute the upper occupancy bound $u_t(s_{t,k}, a_{t,k}) = \max_{\hat{P} \in \mathcal{C}_t} q^{\hat{P}, \pi_t}(s_{t,k}, a_{t,k})$.
- Compute the loss estimator $\hat{\ell}_t(s_{t,k}, a_{t,k})$ by Equation (4).
- Update transition counts for each $k$,
- $N_{t+1}(s_{t+1,k}, a_{t+1,k}, s_{t+1,k+1}) \leftarrow N_t(s_{t,k}, a_{t,k}, s_{t,k+1}) + I_t(s_{t,k}, a_{t,k}, s_{t,k+1})$,
- $N_{t+1}(s_{t+1,k}, a_{t+1,k}) \leftarrow N_t(s_{t,k}, a_{t,k}) + I_t(s_{t,k}, a_{t,k})$.
- Construct the empirical confidence set $\bar{\mathcal{C}}_{t+1}$.
- Construct the likelihood-calibrated confidence set $\widehat{\mathcal{C}}_{t+1}$.
- Update the nested confidence set, $\mathcal{C}_{t+1} = \mathcal{C}_t \cap \bar{\mathcal{C}}_{t+1} \cap \widehat{\mathcal{C}}_{t+1}$.
- **if** $\mathcal{C}_{t+1} = \varnothing$ **then**
  - Execute an arbitrary Markov policy in all remaining episodes.
  - **break**
- **end**
- Compute the FTRL occupancy $\hat{q}_{t+1}$ using Equation (2).
- Compute the policy $\pi_{t+1}$ using Equation (3).

**end**

## 4.1 EMPIRICAL TRANSITION ESTIMATION

The *empirical transition* is a widely applied and important estimator in adversarial RL with unknown transitions. For each episode $t > 1$, let $N_t(s, a)$ and $N_t(s, a, s')$ be the total numbers of visited state-action pairs $(s, a)$ and triples $(s, a, s')$, respectively, before episode $t$. Clearly, $N_t(s, a) = \sum_{s' \in \mathcal{S}_{k+1}} N_t(s, a, s')$, for any $s \in \mathcal{S}_k$ and $a \in \mathcal{A}$. The empirical transition estimate is

$$\bar{P}_t(s' \mid s, a) = \begin{cases} \dfrac{N_t(s, a, s')}{N_t(s, a)}, & N_t(s, a) > 0, \\ \dfrac{1}{|S_{k+1}|}, & N_t(s, a) = 0. \end{cases} \tag{1}$$

We retain coordinate-wise empirical Bernstein confidence bounds, as used in prior work (Jin et al., 2020). Define the empirical confidence set for episode $t$ by

$$\bar{\mathcal{C}}_t = \{\hat{P} : |\hat{P}(s'|s,a) - \bar{P}_t(s'|s,a)| \le \bar{\rho}_t(s'|s,a), \forall (s,a,s') \in \mathcal{S}_k \times \mathcal{A} \times \mathcal{S}_{k+1}, k = 0, 1, \ldots, K-1\},$$

where the confidence radius $\bar{\rho}_t(s'|s,a)$ is defined by

$$\bar{\rho}_t(s'|s,a) = \min\{1, \sqrt{\frac{4\bar{P}_t(s'|s,a)\ln(\frac{8|\mathcal{S}||\mathcal{A}|T}{\delta})}{\max\{1, N_t(s,a)-1\}}} + \frac{14\ln(\frac{8|\mathcal{S}||\mathcal{A}|T}{\delta})}{3\max\{1, N_t(s,a)-1\}}\},$$

for a confidence parameter $\delta$. By the empirical Bernstein inequality and union bounds, we have the following coverage property (see Appendix B.1 for the proof).

**Lemma 1.** *With probability at least $1 - \frac{\delta}{2}$, we have $P \in \bar{\mathcal{C}}_t$ for any $t$.*

**Remark 1.** *Empirical confidence sets $\bar{\mathcal{C}}_t$ rely solely on observed transitions and can remain broad for infrequently visited state–action pairs. Since they do not incorporate prior transition predictions, they cannot exploit accurate predictions to obtain tighter uncertainty bounds from the same observations.*

### 4.2 Confidence Sets from Transition Predictions

We use the prior transition prediction $\tilde{P}$ as a reference for likelihood-ratio testing: a candidate kernel is retained unless the observed transitions provide sufficient evidence favoring the prior over it. This construction incorporates prior information while preserving statistical coverage regardless of prediction accuracy, provided that $\tilde{P}$ is fixed before data collection (Emmenegger et al., 2023). For each $(s, a) \in \mathcal{S}_k \times \mathcal{A}$, define $L_t(P'; s, a) = \prod_{\substack{1 \le \tau < t \\ (s_{\tau,k}, a_{\tau,k}) = (s,a)}} P'(s_{\tau,k+1} \mid s, a)$, with empty products equal to 1. We then construct a prior-informed likelihood-ratio confidence set

$$\widehat{\mathcal{C}}_t = \{\hat{P} : L_t(\hat{P}; s, a) \ge \alpha L_t(\tilde{P}; s, a), \forall (s, a, s') \in \mathcal{S}_k \times \mathcal{A} \times \mathcal{S}_{k+1}, k = 0, 1, \ldots, K-1\},$$

where $\alpha = \frac{\delta}{2|\mathcal{S}||\mathcal{A}|}$. We consider all valid next-state distributions, including those assigning positive probability to states that have zero probability under the prior prediction, and retain those satisfying the likelihood threshold.

By Ville's inequality and the union bounds, we have the following coverage property for $\widehat{\mathcal{C}}_t$ (see Appendix B.2 for the proof).

**Lemma 2.** *With probability at least $1 - \frac{\delta}{2}$, we have $P \in \hat{\mathcal{C}}_t$ for any $t$.*

Lemma 2 establishes the coverage of the prior-informed likelihood confidence set $\widehat{\mathcal{C}}_t$. Proposition 1 (proved in Appendix B.3) shows that, under $N_t(s, a) \ge 1$ and $L_t(\tilde{P}; s, a) > 0$, this set is equivalent to a KL region centered at the empirical transition estimate, with radius modified by the KL discrepancy between the empirical estimate and the prior prediction.

**Proposition 1.** *If $N_t(s, a) > 0$ and $L_t(\tilde{P}; s, a) > 0$, the above confidence set $\widehat{\mathcal{C}}_t$ is equivalent to*

$$\widehat{\mathcal{C}}_t = \{\hat{P} : \mathrm{KL}\left(\bar{P}_t(\cdot|s, a) \,\middle\|\, \hat{P}(\cdot|s, a)\right) \le \hat{\rho}_t(s, a), \forall (s, a) \in \mathcal{S}_k \times \mathcal{A}, k = 0, 1, \ldots, K-1\}$$

*where*

$$\hat{\rho}_t(s, a) = \frac{\ln(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s, a)} + \mathrm{KL}\left(\bar{P}_t(\cdot|s, a) \middle\| \tilde{P}(\cdot|s, a)\right).$$

The KL region is centered at the empirical transition estimate $\bar{P}_t(\cdot \mid s, a)$. The full support of the prior prediction $\tilde{P}(\cdot \mid s, a)$ ensures that $\mathrm{KL}\Big(\bar{P}_t(\cdot \mid s, a) \,\|\, \tilde{P}(\cdot \mid s, a)\Big) < \infty$. For a fixed sample count, its radius increases with the KL divergence between the empirical estimate and the prior prediction $\tilde{P}(\cdot \mid s, a)$, yielding a smaller region when they agree more closely. If $N_t(s, a) = 0$, the likelihood constraint is vacuous, and the confidence set equals the entire simplex $\Delta(\mathcal{S})$.

**Remark 2.** *The likelihood-calibrated confidence set $\widehat{\mathcal{C}}_t$ always contains the prior prediction $\tilde{P}$. Thus, an inaccurate prior prevents the set from contracting to the true kernel, even as observations accumulate. Furthermore, if the prior assigns zero probability to an observed transition, its likelihood remains zero at all subsequent episodes, and the corresponding confidence set $\widehat{\mathcal{C}}_t$ remains the entire simplex $\Delta(\mathcal{S})$.*

### 4.3 Nested Confidence Sets

We combine the empirical and likelihood-calibrated confidence sets, retaining the empirical restrictions while allowing transition predictions to further constrain the feasible kernels. To ensure nesting over time, we retain all preceding constraints. For each episode $t$, define a nested confidence set $\mathcal{C}_t = \bigcap_{\tau=1}^{t} [\bar{\mathcal{C}}_\tau \cap \widehat{\mathcal{C}}_\tau]$.

By construction, $\mathcal{C}_{t+1} \subseteq \mathcal{C}_t$ and $\mathcal{C}_t \subseteq \bigcap_{\tau=1}^{t} \bar{\mathcal{C}}_\tau \subseteq \bar{\mathcal{C}}_t$, for each episode $t < T$. Thus, the intersection is never larger than the empirical confidence set, regardless of prediction accuracy. When $\widehat{\mathcal{C}}_t = \Delta(\mathcal{S})$, the update intersects the current empirical confidence set with the sets retained from earlier episodes. We have then have the following coverage property for $\mathcal{C}_t$ (see Appendix B.4 for the proof).

**Proposition 2.** *The confidence sets $(\mathcal{C}_t)_{t=1}^{T}$ are convex and nested. Moreover, with probability at least $1-\delta$, $P \in \mathcal{C}_t$ simultaneously for all $t = 1, \ldots, T$; consequently, all $\mathcal{C}_t$ are nonempty on this event.*

### 4.4 TP-FTRL

We incorporate transition predictions into policy selection and bandit loss estimation through the nested confidence sets $\mathcal{C}_t$. Since the nested feasible regions facilitate the analysis of its regularization penalty, we adopt occupancy-measure FTRL, following (Jin et al., 2022). Let $\Pi$ denote the set of stochastic Markov policies and define $\Omega(\mathcal{C}_t) = \{q^{P,\pi} : P \in \mathcal{C}_t, \pi \in \Pi\}$. These occupancy sets are convex and satisfy $\Omega(\mathcal{C}_{t+1}) \subseteq \Omega(\mathcal{C}_t)$ by Proposition 2.

At the beginning of episode $t$, provided that $\mathcal{C}_t$ is nonempty, the learner computes

$$\hat{q}_t \in \operatorname*{arg\,min}_{q \in \Omega(\mathcal{C}_t)} \left\{ \sum_{\tau=1}^{t-1} \langle q, \hat{\ell}_\tau \rangle + \Psi(q) \right\}, \tag{2}$$

where $\hat{\ell}_\tau$ is the loss estimate from episode $\tau$, $\eta > 0$ is the learning rate, and $\Psi(q) = \frac{1}{\eta} \sum_{k=0}^{K-1} \sum_{s \in \mathcal{S}_k} \sum_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}_{k+1}} q(s,a,s') \ln q(s,a,s')$ is the Shannon entropy regularizer, with $0 \ln 0 = 0$. The learner extracts the policy

$$\pi_t(a \mid s) = \frac{\hat{q}_t(s,a)}{\sum_{b \in \mathcal{A}} \hat{q}_t(s,b)}, \tag{3}$$

whenever the denominator is positive, choosing an arbitrary action distribution otherwise. The learner executes the resulting Markov policy $\pi_t$ throughout episode $t$ and updates it between episodes using the accumulated bandit feedback.

Standard importance weighting would require the true occupancy $q^{P,\pi_t}(s,a)$, which is unavailable because the transition kernel $P$ is unknown. Following Jin et al. (2020; 2022), we instead compute the upper occupancy bound by $u_t(s,a) = \max_{\hat{P} \in \mathcal{C}_t} q^{\hat{P},\pi_t}(s,a)$. After executing $\pi_t$ and observing the visited pairs and their losses, the learner constructs

$$\hat{\ell}_t(s,a) = \frac{\ell_t(s,a) I_t(s,a)}{u_t(s,a) + \gamma}, \tag{4}$$

where $I_t(s,a)$ indicates whether $(s,a)$ was visited in episode $t$, and $\gamma > 0$ is the implicit-exploration parameter. Unvisited pairs receive an estimate of zero. Thus, the transition predictions affect both policy selection and bandit loss estimation through the confidence set $\mathcal{C}_t$.

On the simultaneous-coverage event, $P \in \mathcal{C}_t$ for every episode $t$, and hence $q^{P,\pi} \in \Omega(\mathcal{C}_t)$ for every Markov policy $\pi$, including the best fixed policy in hindsight. Nevertheless, the optimizer $\hat{q}_t$ in Equations (2) may be induced by a candidate transition kernel in $\mathcal{C}_t$ different from $P$. Consequently, after extracting $\pi_t$, the optimized occupancy $\hat{q}_t$ need not coincide with the true occupancy $q^{P,\pi_t}$ generated when $\pi_t$ is executed in the MDP. This discrepancy, together with the bias introduced by the upper-occupancy implicit-exploration estimator, is accounted for in our regret analysis. If $\mathcal{C}_t = \varnothing$, the learner executes an arbitrary Markov policy in all remaining episodes.

**Efficient computation of prediction-informed FTRL.** By Proposition 2, $\mathcal{C}_t$ is nonempty on the simultaneous coverage event. We next show that, whenever $\mathcal{C}_t \neq \varnothing$, the FTRL problem 2 can be solved efficiently [1]. For any $q \in \Omega(\mathcal{C}_t)$ with $q(s,a) > 0$, the corresponding transition probabilities satisfy $P_q(s'|s,a) = \frac{q(s,a,s')}{q(s,a)}$. Accordingly, the empirical confidence constraints can be expressed

[1] The TP-FTRL update admits an exact reformulation as a finite-dimensional convex optimization problem over extended occupancy measures; see Appendix A

as linear inequalities in the occupancy variables. Moreover, by Proposition 1, every nonvacuous likelihood-calibrated constraint indexed by $\tau < t$ admits the equivalent representation

$$\sum_{s' \in \mathcal{S}_{k+1}} q(s,a)\bar{P}_\tau(s'|s,a) \ln \frac{q(s,a)\bar{P}_\tau(s'|s,a)}{q(s,a,s')} \leq \hat{\rho}_\tau(s,a)q(s,a). \tag{5}$$

The left-hand side of equation 5 is the perspective of the relative-entropy function and is convex in $q$. Together with the linear occupancy-flow and empirical-confidence constraints, and the entropic regularizer, the FTRL update is a finite-dimensional convex relative-entropy program. See details in Appendix A.

**Remark 3.** *The entropic regularizer and the likelihood-calibrated constraints admit exponential-cone representations, while the occupancy-flow and empirical-confidence constraints are linear. Hence, the FTRL update admits an equivalent exponential-cone formulation whose size is polynomial in the number of triples $(s, a, s')$ and likelihood-calibrated constraints in $\mathcal{C}_t$, and can be solved using standard conic optimization solvers. Zero-occupancy rows with $q(s, a) = 0$ are handled by the closed extension of the corresponding perspective function. Once $\hat{q}_t$ is obtained, the policy $\pi_t$ is recovered directly through Equations (3).*

## 5 Theoretical Results

In this section, we analyze the regret of our TP-FTRL.

### 5.1 Worst-case Regret Bound

We first establish a worst-case regret guarantee for TP-FTRL in Theorem 1 (proved in Appendix C.2 for the proof). This result provides a baseline for the analysis of discrepancy between the empirical transition estimates and transition predictions in the next subsection.

**Theorem 1** (Prediction-Independent Regret Bound)**.** *Under the event of Lemma 1, with probability at least $1 - \frac{11\delta}{2}$, TP-FTRL with $\eta = \gamma = \sqrt{\frac{2K \ln\left(\frac{|A|^2|\mathcal{S}|^3}{\delta}\right)}{3|A||\mathcal{S}|T}}$ ensures*

$$Reg_T(\pi^*) = \mathcal{O}\left(K|\mathcal{S}|\sqrt{|\mathcal{A}|T \ln(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta})}\right) = \tilde{\mathcal{O}}(K|\mathcal{S}|\sqrt{|\mathcal{A}|T}).$$

For fixed problem parameters and confidence level, Theorem 1 establishes a regret bound that grows sublinearly with $T$. The guarantee is independent of the accuracy of the prior transition prediction and thus remains valid under arbitrary misspecification. We next examine how the regret guarantee depends on the discrepancy between the prior transition prediction and the empirical transition estimate.

### 5.2 Regret Bounds via Empirical Transition-Prediction Discrepancy

we characterize the regret of TP-FTRL through the discrepancy between the prior transition prediction $\tilde{P}$ and the empirical transition estimate $\bar{P}_t$. To quantify this discrepancy, we introduce a normalized, weighted KL divergence between two probability kernels. We then specialize this definition to the empirical transition estimate and the prior transition prediction.

**Definition 1** (Normalized Discrepancy Ratio)**.** *Let $P_1$ and $P_2$ be probability kernels on the same input and output spaces. Given a positive weight function $\omega$ and $\alpha \in (0, 1)$, define their normalized discrepancy ratio at $x$ by*

$$\mathcal{D}_{w,\alpha}(P_1 \| P_2)(x) = \frac{\omega(x)}{\ln(1/\alpha)} \mathrm{KL}(P_1(\cdot \mid x) \,\|\, P_2(\cdot \mid x)) \,.$$

The normalized discrepancy ratio measures the KL discrepancy between two kernels, weighted by $\omega(x)$ and normalized by $\ln(1/\alpha)$. For each state-action pair $(s, a)$ with $N_t(s, a) \geq 1$, our normalized discrepancy ratio between the empirical transition estimate and the transition prediction is

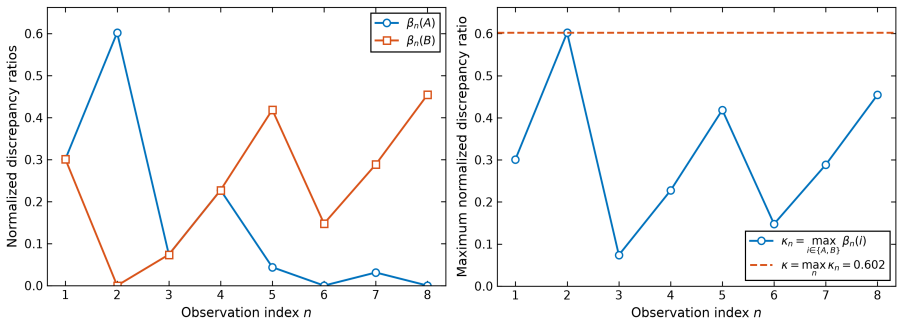
$$\mathcal{D}_{N_t,\alpha}(\bar{P}_t \| \tilde{P})(s,a) = \frac{N_t(s,a)}{\ln(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta})} \mathrm{KL}\left(\bar{P}_t(\cdot \mid s,a) \,\|\, \tilde{P}(\cdot \mid s,a)\right),$$

where $\bar{P}_t$ is the empirical transition estimate, and $\tilde{P}$ is the prior transition prediction. The normalized discrepancy ratio jointly captures the amount of transition data collected at $(s, a)$ and the discrepancy between the empirical transition estimate and the prior transition prediction.

**Definition 2** (Maximum Normalized Discrepancy Ratio)**.** *Under the notation of Definition 1, let $\mathcal{X}$ be the common finite input space. Define the maximum normalized discrepancy ratio by*

$$\kappa = \max_{x \in \mathcal{X}} \mathcal{D}_{\omega,\alpha}(P_1 \| P_2)(x).$$

**Illustrative example.** Consider two state-action pairs $A$ and $B$, each with two possible successor states, and let the supplied transition prediction be $\widetilde{P} = (0.5, 0.5)$ for both pairs. We set $\alpha = 0.1$ and observe one transition from each pair at each observation index $n$. For pair $A$, the observed successors are $1, 1, 0, 1, 0, 0, 1, 0$, while for pair $B$ they are $0, 1, 0, 0, 0, 1, 0, 0$. At each $n$, the empirical transition distribution is constructed from the first $n$ observations, and the normalized discrepancy ratio is $\beta_n(i) = \frac{n}{\ln 10} \mathrm{KL}\big(\bar{P}_n(\cdot \mid i) \,\big\|\, (0.5, 0.5)\big)$ , $i \in \{A, B\}$. Figure 1 illustrates the evolution of these discrepancy ratios. The pair-specific ratios $\beta_n(A)$ and $\beta_n(B)$, and hence their maximum $\kappa_n$, need not vary monotonically as additional transition observations are collected. The overall quantity $\kappa = \max_n \kappa_n$ records the largest normalized discrepancy over the full observation sequence.




(a) Pair-specific normalized discrepancy ratios (b) Maximum normalized discrepancy ratio

Figure 1: Evolution of normalized transition-prediction discrepancy. (a) Pair-specific normalized discrepancy ratios $\beta_n(A)$ and $\beta_n(B)$ as transition observations accumulate. (b) Their pointwise maximum $\kappa_n = \max_{i \in \{A,B\}} \beta_n(i)$; the dashed line denotes the overall maximum $\kappa = \max_n \kappa_n$.

Our maximum normalized discrepancy ratio between the empirical transition estimate and the transition prediction is then defined as $\kappa = \max_t \kappa_t$, where $\kappa_t = \max_{s,a:N_t(s,a)\geq 1} \mathcal{D}_{N_t,\alpha}(\bar{P}_t \| \tilde{P})(s, a)$. Thus, $\kappa_t$ measures the largest normalized transition-prediction discrepancy among all state-action pairs observed before episode $t$, whereas $\kappa$ uniformly summarizes this discrepancy over the full sequence of $T$ episodes.

For a likelihood-calibrated confidence set $\widehat{\mathcal{C}}_t$, its radius satisfies

$$\hat{\rho}_t(s, a) = \frac{\ln(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s, a)} \left(1 + \mathcal{D}_{N_t,\alpha}(\bar{P}_t \| \tilde{P})(s, a)\right) \leq \frac{\ln(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s, a)} (1 + \kappa).$$

Hence, $1 + \kappa$ uniformly controls the inflation of the likelihood-calibrated confidence radius relative to the zero-discrepancy case. The sequence $\{\kappa_t\}_{t=1}^T$ need not be monotone, since the empirical transition estimates evolve as additional transitions are observed. In contrast, $\kappa$ provides a single horizon-wide discrepancy measure used in the regret analysis.

We first consider the idealized zero-discrepancy case $\kappa = 0$, in which the prior transition prediction agrees with the empirical transition estimate at every visited state-action pair throughout learning. the discrepancy term in the likelihood-calibrated confidence radius vanishes, yielding the following improved regret guarantee (see Appendix C.3 for the proof).

**Theorem 2** (Zero-Discrepancy Regret Bound)**.** *Suppose $\kappa = 0$. Under the event of Lemma 2, with probability at least $1 - \frac{11\delta}{2}$, TP-FTRL with $\eta = \gamma = \sqrt{\frac{2K \ln\left(\frac{|A|^2|\mathcal{S}|^3}{\delta}\right)}{3|A||\mathcal{S}|T}}$ ensures*

$$Reg_T(\pi^*) = \mathcal{O}\left(K^{\frac{3}{2}}\sqrt{|\mathcal{S}||\mathcal{A}|T\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}\right) = \tilde{\mathcal{O}}\left(K^{\frac{3}{2}}\sqrt{|\mathcal{S}||\mathcal{A}|T}\right).$$

We next allow a nonzero discrepancy between the prior transition prediction and the empirical transition estimate. The following theorem extends Theorem 2 by quantifying the effect of the maximum normalized discrepancy ratio $\kappa$ (see Appendix C.4 for the proof). Relative to the zero-discrepancy case, the leading regret term deteriorates by only a factor of $\sqrt{1+\kappa}$.

**Theorem 3** (Discrepancy-Dependent Regret Bound)**.** *Under the event of Lemma 2, with probability at least $1 - \frac{11\delta}{2}$, TP-FTRL with $\eta = \gamma = \sqrt{\frac{2K \ln\left(\frac{|A|^2|\mathcal{S}|^3}{\delta}\right)}{3|A||\mathcal{S}|T}}$ ensures*

$$Reg_T(\pi^*) = \mathcal{O}\left(K^{\frac{3}{2}}\sqrt{(1+\kappa)|\mathcal{S}||\mathcal{A}|T\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}\right) = \tilde{\mathcal{O}}\left(K^{\frac{3}{2}}\sqrt{(1+\kappa)|\mathcal{S}||\mathcal{A}|T}\right).$$

The discrepancy-dependent bound in Theorem 3 becomes looser as $\kappa$ increases. However, the prediction-independent worst-case guarantee of Theorem 1 remains valid regardless of the quality of the supplied prediction. Combining these two guarantees yields the following bound (see Appendix C.5 for the proof).

**Theorem 4** (Unified Regret Bound)**.** *Under the event of Proposition 2, with probability at least $1 - 6\delta$, TP-FTRL with $\eta = \gamma = \sqrt{\frac{2K \ln\left(\frac{|A|^2|\mathcal{S}|^3}{\delta}\right)}{3|A||\mathcal{S}|T}}$ ensures*

$$\begin{aligned} Reg_T(\pi^*) &= \mathcal{O}\left(\min\left\{K^{\frac{3}{2}}\sqrt{(1+\kappa)|\mathcal{S}||\mathcal{A}|T\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})},\ K|\mathcal{S}|\sqrt{|\mathcal{A}|T\ln(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta})}\right\}\right) \\ &= \tilde{\mathcal{O}}\left(K\sqrt{|\mathcal{S}||\mathcal{A}|T}\ \min\left\{\sqrt{K(1+\kappa)},\ \sqrt{|\mathcal{S}|}\right\}\right). \end{aligned}$$

Theorem 4 combines the discrepancy-dependent guarantee of Theorem 3 with the prediction-independent guarantee of Theorem 1 by taking their minimum. When $\kappa$ is small, the discrepancy-dependent bound becomes sharper due to the small discrepancy between the empirical transition estimate and the transition prediction. As $\kappa$ increases, this term deteriorates at the rate $\sqrt{1+\kappa}$, while the prediction-independent term remains unchanged and provides a worst-case safeguard. Ignoring logarithmic factors, the discrepancy-dependent bound is smaller whenever $K(1+\kappa) < |\mathcal{S}|$ or equivalently, $\kappa < \frac{|\mathcal{S}|}{K} - 1$. Thus, the benefit of transition predictions is most pronounced when $\frac{|\mathcal{S}|}{K}$ is large and the discrepancy between the empirical transition estimate and the transition prediction is small.

## 6 Conclusion

We studied episodic adversarial MDPs with unknown transitions and potentially misspecified transition predictions. We proposed TP-FTRL, which combines empirical and likelihood-calibrated confidence sets to exploit transition predictions while preserving high-probability coverage and a prediction-independent worst-case regret bound of order $\tilde{\mathcal{O}}(K|\mathcal{S}|\sqrt{|\mathcal{A}|T})$. When the discrepancy between the empirical transition estimate and the transition prediction is small, TP-FTRL admits a sharper regret guarantee. Future work may extend the framework to adversarial transition dynamics and time-varying transition predictions.

## AI use statement

(This section is **required** and does not count toward the page limit.)

In this work, we used ChatGPT solely to improve the grammar, clarity, and readability of the manuscript. We did not use generative AI for any tasks requiring disclosure under the ICLR 2027 AI policy. The authors reviewed all AI-assisted revisions and verified that they preserved the intended technical meaning. We take full responsibility for the final content of the manuscript.

### ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

This work is theoretical in nature and does not involve human participants or personal data. We do not identify any specific ethical concerns arising directly from the theoretical results presented in this paper.

### REPRODUCIBILITY STATEMENT

The assumptions and formal statements of our theoretical results are provided in Sections 3-4. Complete proofs, together with the supporting lemmas and derivations, are presented in Appendices A-C. These materials provide the details needed to independently verify our results.

## REFERENCES

Yasin Abbasi Yadkori, Peter L Bartlett, Varun Kanade, Yevgeny Seldin, and Csaba Szepesvári. Online learning in markov decision processes with adversarially chosen transition probability distributions. *Advances in neural information processing systems*, 26, 2013.

Rohit Agrawal. Finite-sample concentration of the multinomial in relative entropy. *IEEE Transactions on Information Theory*, 66(10):6297–6302, 2020.

Nicolò Cesa-Bianchi and Francesco Orabona. Online learning algorithms. *Annual review of statistics and its application*, 8(1):165–190, 2021.

Liyu Chen and Haipeng Luo. Finding the stochastic shortest path with low regret: The adversarial cost and unknown transition case. In *International Conference on Machine Learning*, pp. 1651–1660. PMLR, 2021.

Yutian Cheng, Canzhe Zhao, and Shuai Li. Policy optimization for adversarial linear mixture mdps with unknown transitions and bandit feedback. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, pp. 1249–1270, 2026.

Yan Dai, Haipeng Luo, Chen-Yu Wei, and Julian Zimmert. Refined regret for adversarial mdps with linear function approximation. In *International Conference on Machine Learning*, pp. 6726–6759. PMLR, 2023.

Travis Dick, Andras Gyorgy, and Csaba Szepesvari. Online learning in markov decision processes with changing cost sequences. In *International Conference on Machine Learning*, pp. 512–520. PMLR, 2014.

Nicolas Emmenegger, Mojmir Mutny, and Andreas Krause. Likelihood ratio confidence sets for sequential decision making. *Advances in Neural Information Processing Systems*, 36:26686–26698, 2023.

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized markov decision processes. In *Conference on learning theory*, pp. 861–898. PMLR, 2015.

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

Tiancheng Jin, Longbo Huang, and Haipeng Luo. The best of both worlds: stochastic and adversarial episodic mdps with unknown transition. *Advances in Neural Information Processing Systems*, 34: 20491–20502, 2021.

Tiancheng Jin, Tal Lancewicki, Haipeng Luo, Yishay Mansour, and Aviv Rosenberg. Near-optimal regret for adversarial mdp with delayed bandit feedback. *Advances in Neural Information Processing Systems*, 35:33469–33481, 2022.

Johannes Kirschner, Andreas Krause, Michele Meziu, and Mojmir Mutny. Confidence estimation via sequential likelihood mixing. *arXiv preprint arXiv:2502.14689*, 2025.

Fang Kong, Xiangcheng Zhang, Baoxiang Wang, and Shuai Li. Improved regret bounds for linear adversarial mdps via linear optimization. *arXiv preprint arXiv:2302.06834*, 2023.

Tal Lancewicki, Aviv Rosenberg, and Dmitry Sotnikov. Delay-adapted policy optimization and improved regret for adversarial mdp with delayed bandit feedback. In *International Conference on Machine Learning*, pp. 18482–18534. PMLR, 2023.

Long-Fei Li, Peng Zhao, and Zhi-Hua Zhou. Dynamic regret of adversarial mdps with unknown transition and linear function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13572–13580, 2024a.

Long-Fei Li, Peng Zhao, and Zhi-Hua Zhou. Improved algorithm for adversarial linear mixture mdps with bandit feedback and unknown transition. In *International Conference on Artificial Intelligence and Statistics*, pp. 3061–3069. PMLR, 2024b.

Yingru Li and Zhiquan Luo. Prior-dependent analysis of posterior sampling reinforcement learning with function approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 559–567. PMLR, 2024.

Haolin Liu, Zakaria Mhammedi, Chen-Yu Wei, and Julian Zimmert. Beating adversarial low-rank mdps with unknown transition and bandit feedback. *Advances in Neural Information Processing Systems*, 37:134645–134700, 2024a.

Haolin Liu, Chen-Yu Wei, and Julian Zimmert. Towards optimal regret in adversarial linear mdps with bandit feedback. In *International Conference on Learning Representations*, volume 2024, pp. 47042–47082, 2024b.

Chenbei Lu, Zaiwei Chen, Tongxin Li, Chenye Wu, and Adam Wierman. Reinforcement learning with imperfect transition predictions: A bellman-jensen approach. *Advances in Neural Information Processing Systems*, 38:102233–102273, 2026.

Lixing Lyu, Jiashuo Jiang, and Wang Chi Cheung. Efficiently solving discounted mdps with predictions on transition matrices. *arXiv preprint arXiv:2502.15345*, 2025.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

H Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *Journal of Machine Learning Research*, 18(90):1–50, 2017.

Ali Nouri. A bayesian sampling approach to exploration in reinforcement learning. In *Proceedings of the...*, 2009.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, pp. 5478–5486. PMLR, 2019a.

Aviv Rosenberg and Yishay Mansour. Online stochastic shortest path with bandit feedback and unknown transition function. *Advances in Neural Information Processing Systems*, 32, 2019b.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2025.

Richard S Sutton, Andrew G Barto, and Andrew Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Andrea Tirinzoni, Riccardo Poiani, and Marcello Restelli. Sequential transfer in reinforcement learning with a generative model. In *International Conference on Machine Learning*, pp. 9481–9492. PMLR, 2020.

Yudan Wang, Kaiyi Ji, Ming Shi, and Shaofeng Zou. Minimax optimal adversarial reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Canzhe Zhao, Ruofeng Yang, Baoxiang Wang, Xuezhou Zhang, and Shuai Li. Learning adversarial low-rank markov decision processes with unknown transition and full-information feedback. *Advances in Neural Information Processing Systems*, 36:59107–59123, 2023.

Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. *Advances in neural information processing systems*, 26, 2013.

# A Efficient Computation of the FTRL Update

This section provides the details underlying Remark 3. By Proposition 2, on the simultaneous coverage event, $\mathcal{C}_t$ is nonempty for every episode $t$. We fix an episode $t$ such that $\mathcal{C}_t \neq \varnothing$ and derive an equivalent exponential-cone formulation of the FTRL update.

Recall that $q(s,a,s')$ denotes the occupancy for a triple $(s,a,s')$ and the occupancy-flow constraints

$$q(s,a) = \sum_{s' \in \mathcal{S}_{k+1}} q(s,a,s'), \qquad s \in \mathcal{S}_k.$$

Whenever $q(s,a) > 0$, the transition probabilities induced by $q$ are given by

$$\hat{P}_q(s'|s,a) = \frac{q(s,a,s')}{q(s,a)}. \tag{6}$$

**Empirical confidence constraints.** For each $(s,a,s')$, define the lower and upper bounds obtained from the empirical confidence sets retained up to episode $t$ by

$$\underline{p}_t(s'|s,a) = \max\left\{0,\ \max_{\tau \le t}\left[\bar{P}_\tau(s'|s,a) - \bar{\rho}_\tau(s'|s,a)\right]\right\},$$

and

$$\overline{p}_t(s'|s,a) = \min\left\{1,\ \min_{\tau \le t}\left[\bar{P}_\tau(s'|s,a) + \bar{\rho}_\tau(s'|s,a)\right]\right\}.$$

Using Equation (6), the empirical confidence constraints are equivalently represented as

$$\underline{p}_t(s'|s,a)q(s,a) \le q(s,a,s') \le \overline{p}_t(s'|s,a)q(s,a), \tag{7}$$

for every triple $(s,a,s')$. Hence, all empirical transition restrictions are linear in the occupancy variables.

**Likelihood-calibrated constraints.** Fix the current episode $t$. For each $\tau \le t$, consider a nonvacuous likelihood-calibrated constraint defining $\mathcal{C}_t$. By Proposition 1, it can be written as

$$\mathrm{KL}\Big(\bar{P}_\tau(\cdot|s,a)\ \Big\|\ \hat{P}(\cdot|s,a)\Big) \le \hat{\rho}_\tau(s,a). \tag{8}$$

For $q(s,a) > 0$, substituting Equation (6) into Equation (8) and multiplying by $q(s,a)$ gives

$$\begin{aligned} &q(s,a)\mathrm{KL}\left(\bar{P}_\tau(\cdot|s,a)\ \middle\|\ \frac{q(s,a,\cdot)}{q(s,a)}\right) \\ &\quad= \sum_{s' \in \mathcal{S}_{k+1}} q(s,a)\bar{P}_\tau(s'|s,a)\ln\left(\frac{q(s,a)\bar{P}_\tau(s'|s,a)}{q(s,a,s')}\right) \le \hat{\rho}_\tau(s,a)q(s,a). \end{aligned} \tag{9}$$

The left-hand side of Equation (9) is convex in the occupancy variables, by the proof of proposition 2. At the boundary, we use the closed extension

$$0\ln\frac{0}{y} = 0 \quad (y \ge 0), \qquad x\ln\frac{x}{0} = +\infty \quad (x > 0).$$

If $q(s,a) = 0$, then nonnegativity and $q(s,a) = \sum_{s'} q(s,a,s')$ imply $q(s,a,s') = 0$ for every $s'$. Hence, the left-hand side of Equation (9) equals zero under the closed extension, and the corresponding occupancy-space constraint is satisfied automatically.

Recall the original FTRL problem

$$\hat{q}_t \in \operatorname*{arg\,min}_{q \in \Omega(\mathcal{C}_t)} \left\{\sum_{\tau=1}^{t-1}\langle q, \hat{\ell}_\tau\rangle + \Psi(q)\right\}, \tag{10}$$

where $\hat{\ell}_\tau$ is the loss estimate from episode $\tau$, $\eta > 0$ is the learning rate, and $\Psi(q) = \frac{1}{\eta}\sum_{k=0}^{K-1}\sum_{s\in\mathcal{S}_k}\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}_{k+1}} q(s,a,s')\ln q(s,a,s')$ is the Shannon entropy regularizer, with $0\log 0 = 0$.

Let $\mathcal{Q}_t$ denote the set of nonnegative occupancy measures satisfying the standard occupancy-flow constraints, Equation (7), and Equation (9) for all $(\tau,s,a)$, where $\tau < t$. We have the following theorem for the relationship between $\mathcal{Q}_t$ and $\mathcal{C}_t$.

**Theorem 5.** *For any episode $t \in [T]$ such that $\mathcal{C}_t \neq \varnothing$, the occupancy-space feasible set $\mathcal{Q}_t$ satisfies*

$$\mathcal{Q}_t = \Omega(\mathcal{C}_t).$$

*Proof.* We first show that $\Omega(\mathcal{C}_t) \subseteq \mathcal{Q}_t$. Let $q = q^{\hat{P},\pi} \in \Omega(\mathcal{C}_t)$ for some $\hat{P} \in \mathcal{C}_t$ and $\pi \in \Pi$. By the definition of the occupancy measure, $q(s,a,s') = q(s,a)\hat{P}(s'|s,a)$. Therefore, the occupancy-flow constraints are satisfied. Moreover, each empirical confidence constraint on $\hat{P}(s'|s,a)$ implies the corresponding linear occupancy constraint after multiplication by $q(s,a)$. Similarly, each nonvacuous likelihood-calibrated constraint implies equation 9 after multiplication by $q(s,a)$. Hence, $q$ satisfies all constraints defining $\mathcal{Q}_t$, and therefore $q \in \mathcal{Q}_t$.

We next show that $\mathcal{Q}_t \subseteq \Omega(\mathcal{C}_t)$. Let $q \in \mathcal{Q}_t$. Since $\mathcal{C}_t \neq \varnothing$ by Proposition 2, fix an arbitrary transition kernel $P' \in \mathcal{C}_t$. We construct a transition kernel $\hat{P}_q$ as follows:

$$\hat{P}_q(s'|s,a) = \begin{cases} \dfrac{q(s,a,s')}{q(s,a)}, & q(s,a) > 0, \\ P'(s'|s,a), & q(s,a) = 0. \end{cases} \tag{11}$$

For $q(s,a) > 0$, nonnegativity of the occupancy variables and $q(s,a) = \sum_{s'} q(s,a,s')$ imply that $\hat{P}_q(\cdot|s,a)$ is a probability distribution. Consider first a state–action pair $(s,a)$ with $q(s,a) > 0$. Dividing the empirical occupancy constraints by $q(s,a)$ shows that $\hat{P}_q(s'|s,a)$ satisfies all empirical confidence constraints associated with $(s,a)$. Likewise, dividing Equation (9) by $q(s,a)$ gives

$$\mathrm{KL}\Big(\bar{P}_\tau(\cdot|s,a) \,\Big\|\, \hat{P}_q(\cdot|s,a)\Big) \leq \hat{\rho}_\tau(s,a),$$

for every nonvacuous likelihood-calibrated constraint associated with $(s,a)$. Thus, $\hat{P}_q(\cdot|s,a)$ satisfies all confidence constraints defining $\mathcal{C}_t$ at this state-action pair.

If $q(s,a) = 0$, we set $\hat{P}_q(\cdot|s,a) = P'(\cdot|s,a)$. Hence, the transition row $\hat{P}_q(\cdot|s,a)$ automatically satisfies all confidence constraints associated with $(s,a)$.

Finally, define the Markov policy

$$\pi_q(a|s) = \frac{q(s,a)}{\sum_{b\in\mathcal{A}} q(s,b)},$$

whenever the denominator is positive, and choose an arbitrary action distribution otherwise. Together with the occupancy-flow constraints, the definitions of $\hat{P}_q$ and $\pi_q$ imply that $q = q^{\hat{P}_q,\pi_q}$. Since $\hat{P}_q \in \mathcal{C}_t$ and $\pi_q \in \Pi$, it follows that $q \in \Omega(\mathcal{C}_t)$. Therefore, the result follows. □

By Theorem 5, the feasible set $\Omega(\mathcal{C}_t)$ can be equivalently represented by $\mathcal{Q}_t$. Hence, the FTRL update can be reformulated directly over the occupancy variables as follows.

**Occupancy-space FTRL formulation.** We now show the occupancy-space FTRL formulation.

$$\begin{aligned}
\min_{q\geq 0} \quad & \sum_{k=0}^{K-1} \sum_{s\in\mathcal{S}_k} \sum_{a\in\mathcal{A}} \sum_{\tau=1}^{t-1} \hat{\ell}_\tau(s,a) q(s,a) + \frac{1}{\eta} \sum_{k=0}^{K-1} \sum_{s\in\mathcal{S}_k} \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}_{k+1}} q(s,a,s') \ln q(s,a,s') \qquad (12) \\
\text{s.t.} \quad & q(s,a) = \sum_{s'\in\mathcal{S}_{k+1}} q(s,a,s') \\
& \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}_{k-1}} q(s,a,s') = \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}_{k+1}} q(s,a,s'), \\
& \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}_1} q(s_0,a,s') = 1, \\
& \underline{p}_t(s'|s,a) q(s,a) \leq q(s,a,s') \leq \overline{p}_t(s'|s,a) q(s,a), \\
& \sum_{s'\in\mathcal{S}_{k+1}} q(s,a) \bar{P}_\tau(s'|s,a) \ln\left(\frac{q(s,a)\bar{P}_\tau(s'|s,a)}{q(s,a,s')}\right) \leq \hat{\rho}_\tau(s,a) q(s,a). \qquad \tau < t
\end{aligned}$$

The objective in the formulation 12 is convex, the occupancy-flow and empirical-confidence constraints are linear, and the likelihood-calibrated constraints are convex. Thus the above formulation is a finite-dimensional convex optimization problem.

**Exponential-cone representation.** We next show that the occupancy-space FTRL formulation (12) admits an equivalent exponential-cone representation. Let

$$\mathcal{K}_{\exp} = \left\{(x, y, z) \in \mathbb{R}^3 : y > 0,\ y \exp(x/y) \le z\right\}$$

denote the closed exponential cone. For $u, v \ge 0$, the relation

$$r \ge u \ln \frac{u}{v}$$

is equivalently represented as

$$(-r, u, v) \in \mathcal{K}_{\exp}. \tag{13}$$

For each $(s, a, s')$, introduce an auxiliary variable $h(s, a, s')$. Since

$$q(s, a, s') \ln q(s, a, s') = q(s, a, s') \ln \frac{q(s, a, s')}{1},$$

the entropic term can be represented by

$$(-h(s, a, s'),\ q(s, a, s'),\ 1) \in \mathcal{K}_{\exp}, \tag{14}$$

which is equivalent to

$$h(s, a, s') \ge q(s, a, s') \ln q(s, a, s').$$

Hence, the entropic regularizer can be replaced by the linear term

$$\frac{1}{\eta} \sum_{k=0}^{K-1} \sum_{s \in \mathcal{S}_k} \sum_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}_{k+1}} h(s, a, s').$$

Next, consider a nonvacuous likelihood-calibrated constraint associated with $(\tau, s, a)$, $\tau < t$. For each $s' \in \mathcal{S}_{k+1}$, introduce an auxiliary variable $r_\tau(s, a, s')$ and impose

$$\left(-r_\tau(s, a, s'),\ q(s, a)\bar{P}_\tau(s'|s, a),\ q(s, a, s')\right) \in \mathcal{K}_{\exp}. \tag{15}$$

By equation 13, this is equivalent to

$$r_\tau(s, a, s') \ge q(s, a)\bar{P}_\tau(s'|s, a) \ln \frac{q(s, a)\bar{P}_\tau(s'|s, a)}{q(s, a, s')}.$$

Therefore, the likelihood-calibrated constraint

$$\sum_{s' \in \mathcal{S}_{k+1}} q(s, a)\bar{P}_\tau(s'|s, a) \ln \frac{q(s, a)\bar{P}_\tau(s'|s, a)}{q(s, a, s')} \le \hat{\rho}_\tau(s, a) q(s, a)$$

is equivalently represented by equation 15 together with

$$\sum_{s' \in \mathcal{S}_{k+1}} r_\tau(s, a, s') \le \hat{\rho}_\tau(s, a) q(s, a), \qquad \tau < t. \tag{16}$$

Consequently, the FTRL update can be equivalently written as

$$\begin{aligned}
\min_{q,h,r} \quad & \sum_{k=0}^{K-1} \sum_{s \in \mathcal{S}_k} \sum_{a \in \mathcal{A}} \sum_{\tau=1}^{t-1} \hat{\ell}_\tau(s, a) q(s, a) + \frac{1}{\eta} \sum_{k=0}^{K-1} \sum_{s \in \mathcal{S}_k} \sum_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}_{k+1}} h(s, a, s') \\
\text{s.t.} \quad & q(s, a) = \sum_{s' \in \mathcal{S}_{k+1}} q(s, a, s'), \\
& \underline{p}_t(s'|s, a) q(s, a) \le q(s, a, s') \le \overline{p}_t(s'|s, a) q(s, a), \\
& (-h(s, a, s'),\ q(s, a, s'),\ 1) \in \mathcal{K}_{\exp}, \\
& \left(-r_\tau(s, a, s'),\ q(s, a)\bar{P}_\tau(s'|s, a),\ q(s, a, s')\right) \in \mathcal{K}_{\exp}, \\
& \sum_{s' \in \mathcal{S}_{k+1}} r_\tau(s, a, s') \le \hat{\rho}_\tau(s, a) q(s, a). \qquad \tau < t
\end{aligned} \tag{17}$$

The objective in equation 17 is linear, and all constraints are either linear or exponential-cone constraints. Therefore, the occupancy-space FTRL problem admits an equivalent exponential-cone formulation. In particular, it requires one exponential-cone block for each triple $(s, a, s')$ in the entropic regularizer and one additional block for each $(\tau, s, a, s')$ associated with a nonvacuous likelihood-calibrated constraint. Hence, the size of the resulting formulation is polynomial in the number of triples $(s, a, s')$ and the number of likelihood-calibrated constraints defining $\mathcal{C}_t$.

# B Proofs for Confidence sets

## B.1 Proof of Lemma 1

By the empirical Bernstein inequality (Maurer & Pontil (2009), Theorem 4), we have the following statement.

**Lemma 3.** *Let $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)$ consist of $n \geq 2$ i.i.d. random variables taking values in $[0, 1]$. Then we have, with probability at least $1 - \delta$, that*

$$|\mathbb{E}[Z] - V_n(\mathbf{Z})| \leq \sqrt{\frac{2V_n(\mathbf{Z})\ln(4/\delta)}{n}} + \frac{7\ln(4/\delta)}{3(n-1)},$$

*where $V_n(\mathbf{Z}) = \frac{1}{n(n-1)}\sum_{1\leq i<j\leq n}(Z_i - Z_j)^2$.*

By Lemma 3 and a union bound over all state-action pairs, we have with probability at least $1 - \frac{\delta}{2}$, for all $(s, a, s') \in \mathcal{S}_k \times \mathcal{A} \times \mathcal{S}_{k+1}, k = 0, 1, \ldots, K-1$ and $t$,

$$\left|P(s'|s,a) - \bar{P}_i(s'|s,a)\right| \leq \sqrt{\frac{2\bar{P}_i(s'|s,a)(1-\bar{P}_i(s'|s,a))\ln\left(\frac{8|\mathcal{S}|^2|\mathcal{A}|T}{\delta}\right)}{\max\{1, N_t(s,a)-1\}}} + \frac{7\ln\left(\frac{8|\mathcal{S}|^2|\mathcal{A}|T}{\delta}\right)}{3\max\{1, N_t(s,a)-1\}} \tag{18}$$

$$\leq \sqrt{\frac{4\bar{P}_i(s'|s,a)\ln\left(\frac{8|\mathcal{S}||\mathcal{A}|T}{\delta}\right)}{\max\{1, N_t(s,a)-1\}}} + \frac{14\ln\left(\frac{8|\mathcal{S}||\mathcal{A}|T}{\delta}\right)}{3\max\{1, N_t(s,a)-1\}}. \tag{19}$$

Thus, the result follows.

## B.2 Proof of Lemma 2

For each $(s, a)$, we define the likelihood-ratio process

$$M_t(s,a) = \frac{L_t(\tilde{P}; s, a)}{L_t(P; s, a)},$$

where $P$ is the true transition kernel. Note that $L_t(P; s, a) > 0$ and $M_1(s, a) = 1$ due to no observations before $t = 1$. Then we will show that $M_t(s, a)$ is a nonnegative supermartingale. Let $\mathcal{F}_t$ denote the history before episode $t$; that is, $\mathcal{F}_t$ covers everything observed before episode $t$. Let $\mathcal{G}_{t,k}$ denote the history available immediately before observing the next state at layer $k$ of episode $t$; that is, $\mathcal{G}_{t,k}$ covers everything observed up to the selection of an action but before observing the next state $s_{t,k+1}$. $I_t(s, a)$ denotes whether a state-action pair $(s, a)$ was visited in episode $t$. Clearly, $N_t(s, a) = N_{t-1}(s, a) + I_t(s, a)$.

In episode $t$, for any $(s, a)$, $I_t(s, a)$ is either 0 or 1. If $I_t(s, a) = 0$, $M_{t+1}(s, a) = M_t(s, a)$, and further $\mathbb{E}[M_{t+1}(s, a)|\mathcal{F}_t] = M_t(s, a)$. If $I_t(s, a) = 1$, by the definition of the likelihood-ratio process, we have

$$M_{t+1}(s,a) = M_t(s,a)\frac{\tilde{P}(s_{t,k+1}|s,a)}{P(s_{t,k+1}|s,a)}.$$

Note that the true transition kernel $P$ is unknown to the learner, but the adversary knows it. This means that at layer $k$ of episode $t$, after observing state $s = s_{t,k}$, if the learner selects action $a = a_{t,k}$,

then the next state $s_{t,k+1}$ follows $P(\cdot|s,a)$, and the learner observes $s_{t,k+1}$. Thus, we have

$$\begin{aligned}\mathbb{E}[M_{t+1}(s,a)|\mathcal{G}_{t,k}] &= \sum_{s_{t,k+1}\in\mathcal{S}_{k+1}} M_t(s,a)\frac{\tilde{P}(s_{t,k+1}|s,a)}{P(s_{t,k+1}|s,a)}P(s_{t,k+1}|s,a) \\ &= \sum_{s_{t,k+1}\in\mathcal{S}_{k+1}} M_t(s,a)\tilde{P}(s_{t,k+1}|s,a)\end{aligned} \tag{20}$$

Since $\mathcal{S}_{k+1} \subseteq \mathcal{S}$, we have $\sum_{s_{t,k+1}\in\mathcal{S}_{k+1}} M_t(s,a)\tilde{P}(s_{t,k+1}|s,a) \leq M_t(s,a)$. Thus, $\mathbb{E}[M_{t+1}(s,a)|\mathcal{G}_{t,k}] \leq M_t(s,a)$. Note that $\mathcal{F}_t \subset \mathcal{G}_{t,k}$. By the tower rule of conditional expectations, we have

$$\mathbb{E}[M_{t+1}(s,a)|\mathcal{F}_t] = \mathbb{E}[\mathbb{E}[M_{t+1}(s,a)|\mathcal{G}_{t,k}]|\mathcal{F}_t] \leq M_t(s,a).$$

Therefore, $M_t(s,a)$ is a nonnegative supermartingale. Combined with $M_1(s,a) = 1$, Ville's inequality gives

$$\Pr\left\{\max_{1\leq t\leq T} M_t(s,a) \geq \varepsilon\right\} \leq \frac{1}{\varepsilon}, \qquad \varepsilon > 0.$$

Taking $\varepsilon = 1/\alpha$, we obtain

$$\Pr\left\{\max_{1\leq t\leq T} M_t(s,a) \geq \frac{1}{\alpha}\right\} \leq \alpha.$$

For a state-action pair $(s,a)$, the true transition distribution fails the likelihood threshold at episode $t$ if $L_t(P;s,a) < \alpha L_t(\tilde{P};s,a)$. This means

$$M_t(s,a) = \frac{L_t(\tilde{P};s,a)}{L_t(P;s,a)} > \frac{1}{\alpha}.$$

Consequently, we have

$$\Pr\left\{\exists\, t \leq T : L_t(P;s,a) < \alpha L_t(\tilde{P};s,a)\right\} \leq \Pr\left\{\max_{1\leq t\leq T} M_t(s,a) \geq \frac{1}{\alpha}\right\} \leq \alpha.$$

By definition, the true kernel is excluded from $\widehat{\mathcal{C}}_t$ exactly when its likelihood falls below the threshold for at least one state–action pair. Thus we have

$$\begin{aligned}\Pr\left\{\exists\, t \leq T : P \notin \widehat{\mathcal{C}}_t\right\} &\leq \sum_{k=0}^{K-1}\sum_{s\in\mathcal{S}_k}\sum_{a\in\mathcal{A}} \Pr\left\{\exists\, t \leq T : M_t(s,a) > \frac{1}{\alpha}\right\} \\ &\leq \alpha\sum_{k=0}^{K-1}|\mathcal{S}_k||\mathcal{A}| \leq \alpha|\mathcal{S}||\mathcal{A}| = \frac{\delta}{2}.\end{aligned}$$

The result, hence, follows.

### B.3 Proof of Proposition 1

By the definition of $L_t(\hat{P};s,a)$ for any $(s,a)$ and any candidate transition kernel $\hat{P}$, we have the following equivalent formula.

$$L_t(\hat{P};s,a) = \prod_{s':N_t(s,a,s')>0}\left(\hat{P}(s'|s,a)\right)^{N_t(s,a,s')}.$$

For the empirical transition estimate $\bar{P}$ and prior transition prediction $\tilde{P}$, we also have

$$L_t(\bar{P}_t;s,a) = \prod_{s':N_t(s,a,s')>0}\left(\bar{P}_t(s'|s,a)\right)^{N_t(s,a,s')}, \tag{21}$$

$$L_t(\tilde{P};s,a) = \prod_{s':N_t(s,a,s')>0}\left(\tilde{P}(s'|s,a)\right)^{N_t(s,a,s')}. \tag{22}$$

Note that $\hat{P}(s'|s,a) > 0$ by the construction of the confidence set $\mathcal{C}_t$ for any $(s,a)$ and $t$. Then we have

$$\frac{L_t(\bar{P}_t;s,a)}{L_t(\hat{P};s,a)} = \prod_{s':N_t(s,a,s')>0} \left(\frac{\bar{P}_t(s'|s,a)}{\hat{P}(s'|s,a)}\right)^{N_t(s,a,s')}. \tag{23}$$

After taking logarithms for Equation (23), we have

$$\begin{aligned}\ln\left(\frac{L_t(\bar{P}_t;s,a)}{L_t(\hat{P};s,a)}\right) &= \sum_{s'} N_t(s,a,s')\ln\left(\frac{\bar{P}_t(s'|s,a)}{\hat{P}(s'|s,a)}\right) = \sum_{s'} N_t(s,a)\bar{P}(s'|s,a)\ln\left(\frac{\bar{P}_t(s'|s,a)}{\hat{P}(s'|s,a)}\right) \\ &= N_t(s,a)\mathrm{KL}\left(\bar{P}_t(\cdot|s,a)\,\middle\|\,\hat{P}(\cdot|s,a)\right) \end{aligned} \tag{24}$$

Similarly, we also have

$$\ln\left(\frac{L_t(\bar{P}_t;s,a)}{L_t(\tilde{P};s,a)}\right) = N_t(s,a)\mathrm{KL}\left(\bar{P}_t(\cdot|s,a)\,\middle\|\,\tilde{P}(\cdot|s,a)\right) \tag{25}$$

By the construction of $\mathcal{C}_t$, for any candidate transition kernel $\hat{P}$ and any triple $((s,a,s') \in \mathcal{S}_k \times \mathcal{A} \times \mathcal{S}_{k+1}, k = 0,1,\ldots,K-1$,

$$L_t(\hat{P};s,a) \geq \alpha L_t(\tilde{P};s,a), \tag{26}$$

where $\alpha = \frac{\delta}{2|\mathcal{S}||\mathcal{A}|}$. After taking logarithms of Equation (26), multiplying by -1, and adding $\ln(L_t(\bar{P}_t;s,a))$, we have

$$\begin{aligned}\ln(L_t(\bar{P}_t;s,a)) - \ln(L_t(\hat{P};s,a)) &\leq \ln(\frac{1}{\alpha}) + \ln(L_t(\bar{P}_t;s,a)) - \ln(L_t(\tilde{P}_t;s,a)) \\ \ln\left(\frac{L_t(\bar{P}_t;s,a)}{L_t(\hat{P};s,a)}\right) &\leq \ln(\frac{1}{\alpha}) + \ln\left(\frac{L_t(\bar{P}_t;s,a)}{L_t(\tilde{P};s,a)}\right) \\ N_t(s,a)\mathrm{KL}\left(\bar{P}_t(\cdot|s,a)\,\middle\|\,\hat{P}(\cdot|s,a)\right) &\leq \ln(\frac{1}{\alpha}) + N_t(s,a)\mathrm{KL}\left(\bar{P}_t(\cdot|s,a)\,\middle\|\,\tilde{P}(\cdot|s,a)\right) \end{aligned} \tag{27}$$

Dividing Equation (27) by $N_t(s,a)$ completes the proof.

### B.4 Proof of Proposition 2

By the construction of the nested confidence set $\mathcal{C}_t$, it is clearly nested for each episode $t$. In each episode $t$, we have

$$\begin{aligned}\Pr\{P \in \mathcal{C}_t\} = \Pr\left\{P \in \widehat{\mathcal{C}}_t \cap \bar{\mathcal{C}}_t\right\} &= 1 - \Pr\left\{P \notin \widehat{\mathcal{C}}_t \cup \bar{\mathcal{C}}_t\right\} \\ &\geq 1 - \Pr\left\{P \notin \widehat{\mathcal{C}}_t\right\} - \Pr\left\{P \notin \bar{\mathcal{C}}_t\right\}. \end{aligned}$$

By Lemma 1, Lemma 2, and a union bound, we have $\Pr\left\{P \notin \widehat{\mathcal{C}}_t\right\} < \frac{\delta}{2T}$ and $\Pr\left\{P \notin \bar{\mathcal{C}}_t\right\} < \frac{\delta}{2T}$. Thus, we have

$$\Pr\{P \in \mathcal{C}_t\} \geq 1 - \frac{\delta}{T}.$$

This means $\Pr\{P \notin \mathcal{C}_t\} < \frac{\delta}{T}$. Therefore, $\Pr\left\{P \in \widehat{\mathcal{C}}_t\right\} \geq 1-\delta$. Since every $\mathcal{C}_t$ contains $P$, $\mathcal{C}_t \neq \emptyset$ for all $t \in [T]$.

We now show that $(\mathcal{C}_t)_{t=1}^T$ is convex. Each empirical confidence constraint can be written as

$$\bar{P}_t(s' \mid s,a) - \bar{\rho}_t(s' \mid s,a) \leq \hat{P}(s' \mid s,a) \leq \bar{P}_t(s' \mid s,a) + \bar{\rho}_t(s' \mid s,a).$$

These inequalities are affine in the coordinates of $\hat{P}$. Together with nonnegativity and normalization,

$$\hat{P}(s' \mid s,a) \geq 0, \qquad \sum_{s' \in \mathcal{S}_{k+1}} \hat{P}(s' \mid s,a) = 1,$$

they define a convex set. Hence, $\bar{\mathcal{C}}_t$ is convex for every $t$.

By the preceding KL representation of the likelihood calibrated confidence set $\widehat{\mathcal{C}}_t$ for each episode $t$,

$$\widehat{\mathcal{C}}_t = \left\{ P : \mathrm{KL}\Big(\bar{P}_t(\cdot \mid s,a) \,\Big\|\, \hat{P}(\cdot \mid s,a)\Big) \le \hat{\rho}_t(s,a) \quad \forall s,a \right\}.$$

Here, both $\bar{P}_t(\cdot|s,a)$ and $\hat{\rho}_t(s,a)$ are fixed with respect to the candidate kernel $\hat{P}$. We show that $\widehat{\mathcal{C}}_t$ is convex. For any fixed probability vector $r \in \Delta(\mathcal{S})$,

$$\mathrm{KL}(r\|q) = \sum_{s':\, r(s')>0} r(s') \log r(s') - \sum_{s':\, r(s')>0} r(s') \log q(s').$$

We use the convention that $\mathrm{KL}(r\|q) = +\infty$ if $r(s') > 0$ and $q(s') = 0$ for some $s'$. The first sum is independent of $q$, while each term in the second sum is a nonnegative multiple of the convex function $-\log q(s')$, extended to take the value $+\infty$ at zero. Hence $q \mapsto \mathrm{KL}(r\|q)$ is convex on $\Delta(\mathcal{S})$.

Now take any $P_1, P_2 \in \widehat{\mathcal{C}}_t$ and $\lambda \in (0,1)$, and define

$$P_\lambda(\cdot \mid s,a) = \lambda P_1(\cdot \mid s,a) + (1-\lambda) P_2(\cdot \mid s,a) \qquad \text{for every } (s,a).$$

Moreover, for every $(s,a)$, the convexity of the KL divergence in its second argument gives

$$\begin{aligned}
&\mathrm{KL}\big(\bar{P}_t(\cdot \mid s,a) \,\big\|\, P_\lambda(\cdot \mid s,a)\big) \\
&\quad \le \lambda\, \mathrm{KL}\big(\bar{P}_t(\cdot \mid s,a) \,\big\|\, P_1(\cdot \mid s,a)\big) + (1-\lambda)\, \mathrm{KL}\big(\bar{P}_t(\cdot \mid s,a) \,\big\|\, P_2(\cdot \mid s,a)\big) \\
&\quad \le \lambda \hat{\rho}_t(s,a) + (1-\lambda)\hat{\rho}_t(s,a) = \hat{\rho}_t(s,a).
\end{aligned}$$

Thus, $P_\lambda$ satisfies every nonvacuous likelihood restriction. For any state–action pair whose likelihood restriction is vacuous, the feasible set is the entire probability simplex, which is convex. Therefore $P_\lambda \in \widehat{\mathcal{C}}_t$. The cases $\lambda = 0$ and $\lambda = 1$ are immediate, establishing the convexity of $\widehat{\mathcal{C}}_t$. Therefore, $\mathcal{C}_t = \widehat{\mathcal{C}}_t \cap \bar{\mathcal{C}}_t$ is also convex. It further shows that $(\mathcal{C}_t)_{t=1}^T$ is convex.

# C Proof of Regret analysis

## C.1 Notation and Preliminaries

We first introduce some necessary notations used in our regret analysis. We let $q_t$ denote $q^{P,\pi_t}$ and $q^* \in \arg\min_{q \in \Delta(P)} \sum_{t=1}^T \langle q, \ell_t \rangle$, where $P$ is the true transition kernel.

For any occupancy measure $q$, let $P_q$ denote its corresponding transition kernel shown in Equation (11), and $\pi_q$ is the policy derived from $q$. For any state-action pair $(s,a)$, given that $s_k = s, a_k = a$, we have

$$q(s,a) = \pi_q(a|s) \sum_{\substack{s_1 \in \mathcal{S}_1, \ldots, s_{k-1} \in \mathcal{S}_{k-1} \\ a_0, \ldots, a_{k-1} \in \mathcal{A}}} \prod_{i=0}^{k-1} \pi_q(a_i|s_i) P_q(s_{i+1}|s_i, a_i). \tag{28}$$

We decompose the regret into four terms, following the prior studies (Jin et al., 2020; 2021; 2022):

$$\begin{aligned}
Reg_T(\pi^*) &= \sum_{t=1}^T \langle q^{P,\pi_t} - q^{P,\pi^*}, \ell_t \rangle = \sum_{t=1}^T \langle q_t - q^*, \ell_t \rangle \\
&= \underbrace{\sum_{t=1}^T \langle q_t - \widehat{q}_t, \ell_t \rangle}_{\text{Error reg 1}} + \underbrace{\sum_{t=1}^T \left\langle \widehat{q}_t, \ell_t - \widehat{\ell}_t \right\rangle}_{\text{Bias 1}} + \underbrace{\sum_{t=1}^T \left\langle \widehat{q}_t - q^*, \widehat{\ell}_t \right\rangle}_{\text{Error reg 2}} + \underbrace{\sum_{t=1}^T \left\langle q^*, \widehat{\ell}_t - \ell_t \right\rangle}_{\text{Bias 2}}.
\end{aligned} \tag{29}$$

## C.2 Proof of Theorem 1

As shown in Equation (29), the regret can be decomposed into four terms. Theorem 6 shows the high-probability upper bound of our TP-FTRL algorithm for the empirical confidence sets, which completes the proof of Theorem 1.

**Theorem 6.** *Under the event of Lemma 1, with probability at least* $1-\frac{11\delta}{2}$*, under adversarial bandit feedback loss functions and unknown transitions, TP-FTRL with* $\eta=\gamma=\sqrt{\frac{2K\ln\left(\frac{|A|^2|\mathcal{S}|^3}{\delta}\right)}{3|A||\mathcal{S}|T}}$ *ensures*

$$\begin{aligned}Reg_T(\pi^*) &= \sum_{t=1}^{T}\langle q^{P,\pi_t}-q^{P,\pi^*},\ell_t\rangle \\ &\le \mathcal{O}\Bigg(K|\mathcal{S}|\sqrt{|\mathcal{A}|T\ln(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta})}+\frac{K\ln(|A||\mathcal{S}|^2)}{\eta}+\frac{\eta}{2}|\mathcal{S}||\mathcal{A}|T+\frac{\eta K}{4\gamma}\ln\frac{K}{\delta} \\ &\qquad +\gamma|\mathcal{S}||\mathcal{A}|T+\frac{K}{\gamma}\ln\frac{|\mathcal{S}||\mathcal{A}|}{\delta}\Bigg) \\ &\le \tilde{\mathcal{O}}(K|\mathcal{S}|\sqrt{|\mathcal{A}|T}).\end{aligned}$$

The proof of Theorem 6 is as follows:

**ERROR REG 1 Term:** We first show the upper bound for the regret error between the true transition kernel and any transition kernel from the empirical confidence set.

**Lemma 4.** *Under the event of Lemma 1, with probability at least* $1-\frac{5\delta}{2}$*, TP-FTRL ensures* ERROR REG $1=\mathcal{O}(K|\mathcal{S}|\sqrt{|\mathcal{A}|T\ln(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta})})$.

**Lemma 5** (Lemma 10, Jin et al. (2020))**.** *With probability at least* $1-2\delta$*, we have for all* $k=0,\dots,K-1$*,*

$$\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_k,\,a\in\mathcal{A}}\frac{q_t(s,a)}{\max\{1,N_t(s,a)\}}=\mathcal{O}(|\mathcal{S}_k||A|\ln T+\ln(K/\delta)) \tag{13}$$

*and*

$$\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_k,\,a\in\mathcal{A}}\frac{q_t(s,a)}{\sqrt{\max\{1,N_t(s,a)\}}}=\mathcal{O}\Big(\sqrt{|\mathcal{S}_k||A|T}+|\mathcal{S}_k||A|\ln T+\ln(K/\delta)\Big). \tag{14}$$

**Lemma 6** (Lemma 5, Jin et al. (2020))**.** *With probability at least* $1-6\delta$*, for any collection of transistion functions* $\{P_t^s\}_{s\in\mathcal{S}}$ *such that* $P_t^s\in\mathcal{P}_{i_t}$ *for all* $s$*, we have*

$$\sum_{t=1}^{T}\sum_{s\in\mathcal{S},\,a\in\mathcal{A}}\left|q^{P_t^s,\pi_t}(s,a)-q_t(s,a)\right|=\mathcal{O}\left(K|\mathcal{S}|\sqrt{|\mathcal{A}|T\ln\left(\frac{T|\mathcal{S}||\mathcal{A}|}{\delta}\right)}\right).$$

**Lemma 7.** *Under the event of Lemma 1, with probability at least* $1-\frac{\delta}{2}$*, for all episodes* $t\in[T]$*, all* $\hat{P}\in\bar{\mathcal{C}}_t$*, and* $(s,a)\in\mathcal{S}_k\times\mathcal{A}$*, we have for all* $k=0,\dots,K-1$*,*

$$|\hat{P}(s'|s,a)-P(s'|s,a)|=\mathcal{O}\left(\sqrt{\frac{P(s'|s,a)\ln(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta})}{\max\{1,N_t(s,a)\}}}+\frac{\ln(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta})}{\max\{1,N_t(s,a)\}}\right).$$

*Proof.* Since $P\in\bar{\mathcal{C}}_t$ for any $(s,a)\in\mathcal{S}_k\times\mathcal{A}$ for all $k=0,\dots,K-1$, and $t\in[T]$, we also have

$$\bar{P}(s'|s,a)-P(s'|s,a)\le\min\{1,\sqrt{\frac{4P(s'|s,a)\ln(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta})}{\max\{1,N_t(s,a)\}}}+\frac{14\ln(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta})}{3\max\{1,N_t(s,a)\}}\}.$$

We consider

$$\bar{P}(s'|s,a)-P(s'|s,a)\le\sqrt{\frac{4P(s'|s,a)\ln(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta})}{\max\{1,N_t(s,a)\}}}+\frac{14\ln(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta})}{3\max\{1,N_t(s,a)\}}.$$

Solving this equation of $\bar{P}(s'|s,a)$ and absorbing the coefficient gives us this lemma. □

*Proof of Lemma 4.* We follow the argument of Jin et al. (2020), using Lemma 5 and their confidence set. In their analysis, the two relevant events have failure probabilities at most $2\delta$ and $4\delta$, respectively. In our setting, Lemma 7 has failure probability at most $\delta/2$. Thus, by a union bound, both events hold simultaneously with probability at least $1 - \frac{5\delta}{2}$.

Thus, under the events of both Lemma 5 and Lemma 7, combining them with Lemma 6, we have

$$\sum_{t=1}^{T} \langle q_t - \widehat{q}_t, \ell_t \rangle = \mathcal{O}\left( K|\mathcal{S}|\sqrt{|\mathcal{A}|T \ln\left(\frac{T|\mathcal{S}||\mathcal{A}|}{\delta}\right)}\right).$$

This finished the proof.

□

**ERROR REG 2 Term:** By Lemma 1, we know $q^* \in \Omega(\bar{\mathcal{C}}_t)$. Thus, ERROR REG 2 is controlled by the regret of TP-FTRL.

**Lemma 8.** *Under the event of Lemma 1, with probability at least* $1 - \frac{3\delta}{2}$*, TP-FTRL ensures* ERROR REG $2 = \mathcal{O}\left(\frac{K \ln(|A||\mathcal{S}|^2)}{\eta} + \frac{\eta}{2}|\mathcal{S}||\mathcal{A}|T + \frac{\eta K}{4\gamma} \ln \frac{K}{\delta}\right)$.

**Lemma 9** (Lemma 11, Jin et al. (2020))**.** *For any sequence of functions* $\alpha_1, \ldots, \alpha_T$ *such that* $\alpha_t \in [0, 2\gamma]^{|\mathcal{S}| \times A}$ *is* $\mathcal{F}_t$*-measurable for all* $t$*, we have, with probability at least* $1 - \delta$*,*

$$\sum_{t=1}^{T} \sum_{s \in \mathcal{S},\, a \in A} \alpha_t(s,a) \left( \widehat{\ell}_t(s,a) - \frac{q_t(s,a)}{u_t(s,a)} \ell_t(s,a) \right) \leq K \ln \frac{K}{\delta}.$$

**Lemma 10.** *TP-FTRL updates with* $\hat{q}_1(s,a,s') = \frac{1}{|\mathcal{S}_k||\mathcal{A}||\mathcal{S}_{k+1}|}$ *for any* $(s,a,s') \in \mathcal{S}_k \times \mathcal{A} \times \mathcal{S}_{k+1}$*, where* $k = 0, 1, \ldots, K-1$*, and*

$$\hat{q}_t \in \arg\min_{q \in \Omega(\bar{\mathcal{C}}_t)} \sum_{\tau=1}^{t-1} \langle q, \hat{\ell}_\tau \rangle + \Psi(q),$$

*where* $\Psi(q) = \frac{1}{\eta} \sum_{k=0}^{K-1} \sum_{s \in \mathcal{S}_k} \sum_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}_{k+1}} q(s,a,s') \ln q(s,a,s')$ *ensures*

$$\sum_{t=1}^{T} \left\langle \widehat{q}_t - q, \widehat{\ell}_t \right\rangle \leq \frac{K \ln(|A||\mathcal{S}|^2)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{k=0}^{K-1} \sum_{(s,a) \in \mathcal{S}_k \times \mathcal{A}} \hat{q}_t(s,a) \hat{\ell}_t(s,a)^2,$$

*for any* $q \in \Omega(\bar{\mathcal{C}}_t)$*, as long as* $\hat{\ell}_t(s,a) \geq 0$ *for any* $t, s, a$*.*

*Proof.* We first define

$$\tilde{q}_{t+1}(s,a,s') = \frac{K \hat{q}_t(s,a,s') e^{-\eta \hat{\ell}_t(s,a)}}{\sum_{s,b,s'} \hat{q}_t(s,b,s') e^{-\eta \hat{\ell}_t(s,b)}}.$$

It is easy to verify $\hat{q}_{t+1} = \arg\min_{q \in \Omega(\bar{\mathcal{C}}_t)} \mathrm{KL}(q || \tilde{q}_{t+1})$, and we also have

$$\eta \left\langle \widehat{q}_t - q, \widehat{\ell}_t \right\rangle = \mathrm{KL}(q||\hat{q}_t) - \mathrm{KL}(q||\tilde{q}_{t+1}) + \mathrm{KL}(\hat{q}_t||\tilde{q}_{t+1}).$$

By the KL Pythagorean inequality, we have $\mathrm{KL}(q||\hat{q}_{t+1}) \leq \mathrm{KL}(q||\tilde{q}_{t+1})$. Then, we have

$$\begin{aligned}
\eta \sum_{t=1}^{T} \left\langle \widehat{q}_t - q, \widehat{\ell}_t \right\rangle &\leq \sum_{t=1}^{T} (\mathrm{KL}(q||\hat{q}_t) - \mathrm{KL}(q||\hat{q}_{t+1}) + \mathrm{KL}(\hat{q}_t||\tilde{q}_{t+1})) \\
&= \mathrm{KL}(q||\hat{q}_1) - \mathrm{KL}(q||\hat{q}_{T+1}) + \sum_{t=1}^{T} \mathrm{KL}(\hat{q}_t||\tilde{q}_{t+1})
\end{aligned}$$

For the first two terms, we have

$$\begin{aligned}\mathrm{KL}(q||\hat{q}_1) - \mathrm{KL}(q||\hat{q}_{T+1}) &= \sum_{s,a,s'} q(s,a,s') \ln \frac{\hat{q}_{T+1}(s,a,s')}{\hat{q}_1(s,a,s')} \\ &= \sum_{k=0}^{K-1} \sum_{s\in\mathcal{S}_k} \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}_{k+1}} q(s,a,s') \ln \hat{q}_{T+1}(s,a,s') |\mathcal{S}_k||\mathcal{A}||\mathcal{S}_{k+1}| \\ &\le \sum_{k=0}^{K-1} \ln(|\mathcal{S}_k||\mathcal{A}||\mathcal{S}_{k+1}|) \\ &\le K \ln(|\mathcal{S}|^2|\mathcal{A}|)\end{aligned}$$

Let $w_t = \sum_{s,b,s'} \hat{q}_t(s,b,s') e^{-\eta\hat{\ell}_t(s,b)}$, and thus $\tilde{q}_{t+1}(s,a,s') = \frac{K\hat{q}_t(s,a,s')e^{-\eta\hat{\ell}_t(s,a)}}{w_t}$. Note that $\sum_{s,a,s'} \hat{q}_t(s,a,s') = K$.

$$\mathrm{KL}(\hat{q}_t||\tilde{q}_{t+1}) = \eta \sum_{s,a,s'} \hat{q}_t(s,a,s')\hat{\ell}_t(s,a) + \sum_{s,a,s'} \hat{q}_t(s,a,s') \ln \frac{w_t}{K} = \eta \sum_{s,a,s'} \hat{q}_t(s,a,s')\hat{\ell}_t(s,a) + K \ln \frac{w_t}{K}.$$

Our last step is to bound $\mathrm{KL}(\hat{q}_t||\tilde{q}_{t+1})$. Note that given $\frac{w_t}{K} > 0$, $\ln \frac{u_t}{K} \le \frac{w_t}{K} - 1$. Then we have

$$\begin{aligned}\mathrm{KL}(\hat{q}_t||\tilde{q}_{t+1}) &\le \eta \sum_{s,a,s'} \hat{q}_t(s,a,s')\hat{\ell}_t(s,a) + w_t - K \\ &= \eta \sum_{s,a,s'} \hat{q}_t(s,a,s')\hat{\ell}_t(s,a) + \sum_{s,a,s'} \hat{q}_t(s,a,s')e^{-\eta\hat{\ell}_t(s,a)} - \sum_{s,a,s'} \hat{q}_t(s,a,s') \\ &= \eta \sum_{s,a,s'} \hat{q}_t(s,a,s')\hat{\ell}_t(s,a) + \sum_{s,a,s'} \hat{q}_t(s,a,s')(e^{-\eta\hat{\ell}_t(s,a)} - 1) \\ &= \sum_{s,a,s'} \hat{q}_t(s,a,s')(\eta\hat{\ell}_t(s,a) + e^{-\eta\hat{\ell}_t(s,a)} - 1) \\ &\le \sum_{s,a,s'} \hat{q}_t(s,a,s')(\eta\hat{\ell}_t(s,a) - 1 + 1 - \eta\hat{\ell}_t(s,a) + \frac{\eta^2\hat{\ell}_t^2(s,a)}{2}) \\ &= \frac{\eta^2}{2} \sum_{s,a,s'} \hat{q}_t(s,a,s')\hat{\ell}_t^2(s,a).\end{aligned}$$

Therefore, we finish the proof.

□

*Proof of Lemma 8.* By Lemma 10, we have for any $q \in \Omega(\bar{\mathcal{C}}_t)$,

$$\sum_{t=1}^{T} \left\langle \widehat{q}_t - q, \widehat{\ell}_t \right\rangle \le \frac{K \ln(|A||\mathcal{S}|^2)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{k=0}^{K-1} \sum_{(s,a)\in\mathcal{S}_k\times\mathcal{A}} \hat{q}_t(s,a)\hat{\ell}_t(s,a)^2.$$

Given that $\hat{\ell}_t(s,a) = \frac{\ell_t(s,a)I_t(s,a)}{u_t(s,a)+\gamma}$, $\hat{q}_t(s,a) \le u_t(s,a)$ by the definition of $u_t(s,a)$, and $\ell_t(s,a) \in [0,1]$, we have

$$\hat{q}_t(s,a)\hat{\ell}_t(s,a)^2 \le \hat{q}_t(s,a)\frac{\ell_t(s,a)\hat{\ell}_t(s,a)}{u_t(s,a)+\gamma} \le \hat{q}_t(s,a)\frac{\hat{\ell}_t(s,a)}{u_t(s,a)} \le \hat{\ell}_t(s,a).$$

Applying Lemma 9 with $\alpha_t(s,a) = 2\gamma$, the largest value allowed by the lemma, gives that with probalibility at least $1-\delta$, we have

$$\sum_{t,s,a} \left( \widehat{\ell}_t(s,a) - \frac{q_t(s,a)}{u_t(s,a)}\ell_t(s,a) \right) \le \frac{K}{2\gamma} \ln \frac{K}{\delta}.$$

Since $P \in \bar{\mathcal{C}}_t$, $q_t \in \Omega(\bar{\mathcal{C}}_t)$ under the event of Lemma 1. This means $q_t(s,a) \leq u_t(s,a)$ with probability at least $1 - \frac{\delta}{2}$. Then we have

$$\begin{aligned}\sum_{t,s,a} \hat{q}_t(s,a)\hat{\ell}_t(s,a)^2 &\leq \sum_{t,s,a} \hat{\ell}_t(s,a) \\ &\leq \sum_{t,s,a} \frac{q_t(s,a)}{u_t(s,a)}\ell_t(s,a) + \frac{K}{2\gamma}\ln\frac{K}{\delta} \\ &\leq |\mathcal{S}||\mathcal{A}|T + \frac{K}{2\gamma}\ln\frac{K}{\delta},\end{aligned}$$

with probability at least $1 - \frac{3\delta}{2}$.

Therefore, we have

$$\sum_{t=1}^{T}\left\langle \widehat{q}_t - q, \widehat{\ell}_t \right\rangle \leq \mathcal{O}\left(\frac{K\ln(|A||\mathcal{S}|^2)}{\eta} + \frac{\eta}{2}|\mathcal{S}||\mathcal{A}|T + \frac{\eta K}{4\gamma}\ln\frac{K}{\delta}\right).$$

Since $q^* \in \Omega(\bar{\mathcal{C}}_t)$, taking $q = q^*$ finishes the proof. □

**BIAS 1 Term:** We bound this term by Lemma 11 as follows.

**Lemma 11.** *Under the event of Lemma 1, with probability at least* $1 - \frac{7\delta}{2}$*, TP-FTRL ensures* BIAS $1 = \mathcal{O}\left(K|\mathcal{S}|\sqrt{|\mathcal{A}|T\ln(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta})} + \gamma|\mathcal{S}||\mathcal{A}|T\right)$.

*Proof.* Let $\mathcal{F}_t$ denote the $\sigma$-algebra including the history and the quantities chosen before the first episode $t$, including $\widehat{q}_t$, $u_t$, and $\ell_t$. These quantities are therefore $\mathcal{F}_t$-measurable. The full loss is included only for the analysis; the learner still receives only bandit feedback. Define scale random variables $X_t$ for each $t$ as

$$X_t = \left\langle \hat{q}_t, \hat{\ell}_t \right\rangle = \sum_{s,a} \hat{q}_t(s,a)\hat{\ell}_t(s,a) = \sum_{s,a} \hat{q}_t(s,a)\frac{\ell_t(s,a)I_t(s,a)}{u_t(s,a)+\gamma}.$$

Since $\hat{\ell}_t(s,a) \in [0,1]$ and $\hat{q}_t(s,a) \leq u_t(s,a)$ for any $(s,a)$, we have

$$X_t \leq \sum_{s,a} u_t(s,a)\frac{I_t(s,a)}{u_t(s,a)} \leq \sum_{s,a} I_t(s,a) = K,$$

which implies that $|\mathbb{E}[X_t|\mathcal{F}_t] - X_t| \leq K$. We also have

$$\mathbb{E}[\mathbb{E}[X_t|\mathcal{F}_t] - X_t \mid \mathcal{F}_t] = \mathbb{E}[X_t|\mathcal{F}_t] - \mathbb{E}[X_t|\mathcal{F}_t] = 0.$$

Note that $\mathbb{E}[X_t|\mathcal{F}_t] - X_t$ is just $\left\langle \hat{q}_t, \mathbb{E}[\hat{\ell}_t|\mathcal{F}_t] - \hat{\ell}_t \right\rangle$. By the Azuma-Hoeffding inequality, with probability at least $1-\delta$, we have

$$\sum_{t=1}^{T}\left\langle \hat{q}_t, \mathbb{E}[\hat{\ell}_t|\mathcal{F}_t] - \hat{\ell}_t \right\rangle \leq K\sqrt{2T\ln(\frac{1}{\delta})}.$$

We have

$$\sum_{t=1}^{T}\left\langle \hat{q}_t, \ell_t - \hat{\ell}_t \right\rangle = \sum_{t=1}^{T}\left\langle \hat{q}_t, \ell_t - \mathbb{E}[\hat{\ell}_t|\mathcal{F}_t] \right\rangle + \sum_{t=1}^{T}\left\langle \hat{q}_t, \mathbb{E}[\hat{\ell}_t|\mathcal{F}_t] - \hat{\ell}_t \right\rangle.$$

We now bound $\sum_{t=1}^{T}\left\langle \hat{q}_t, \ell_t - \mathbb{E}[\hat{\ell}_t|\mathcal{F}_t]\right\rangle$. Note that $\mathbb{E}[I_t(s,a)|\mathcal{F}_t] = q_t(s,a)$. We further have

$$\begin{aligned}
\sum_{t=1}^{T}\left\langle \hat{q}_t, \ell_t - \mathbb{E}[\hat{\ell}_t|\mathcal{F}_t]\right\rangle &= \sum_{t=1}^{T}\sum_{s,a}\hat{q}_t(s,a)\left(\ell_t(s,a) - \frac{\ell_t(s,a)}{u_t(s,a)+\gamma}\mathbb{E}[I_t(s,a)|\mathcal{F}_t]\right) \\
&= \sum_{t=1}^{T}\sum_{s,a}\hat{q}_t(s,a)\frac{\ell_t(s,a)}{u_t(s,a)+\gamma}\left(u_t(s,a)+\gamma - q_t(s,a)\right) \\
&\leq \sum_{t=1}^{T}\sum_{s,a}u_t(s,a)\frac{\ell_t(s,a)}{u_t(s,a)}\left(u_t(s,a)+\gamma - q_t(s,a)\right) \\
&\leq \sum_{t=1}^{T}\sum_{s,a}|(u_t(s,a) - q_t(s,a)| + \sum_{t=1}^{T}\sum_{s,a}\gamma \\
&\leq \sum_{t=1}^{T}\sum_{s,a}|(u_t(s,a) - q_t(s,a)| + \gamma|\mathcal{S}||\mathcal{A}|T.
\end{aligned}$$

By Lemma 4 and the definition of $u_t$, with probability at least $1 - \frac{5\delta}{2}$,

$$\sum_{t=1}^{T}\sum_{s,a}|(u_t(s,a) - q_t(s,a)| \leq \mathcal{O}(K|\mathcal{S}|\sqrt{|\mathcal{A}|T\ln(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta})}).$$

Therefore, we have, with probability at least $1 - \frac{7\delta}{2}$,

$$\begin{aligned}
\textsc{Bias 2} &= \mathcal{O}\left(K|\mathcal{S}|\sqrt{|\mathcal{A}|T\ln(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta})} + \gamma|\mathcal{S}||\mathcal{A}|T + K\sqrt{2T\ln(\frac{1}{\delta})}\right) \\
&= \mathcal{O}\left(K|\mathcal{S}|\sqrt{|\mathcal{A}|T\ln(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta})} + \gamma|\mathcal{S}||\mathcal{A}|T\right).
\end{aligned}$$

□

**BIAS 2 Term:** We bound this term by Lemma 12 as follows.

**Lemma 12.** *Under the event of Lemma 1, with probability at least* $1 - \frac{3\delta}{2}$*, TP-FTRL ensures* BIAS $2 = \mathcal{O}\left(\frac{K}{\gamma}\ln\frac{|\mathcal{S}||\mathcal{A}|}{\delta}\right)$.

*Proof.* Fix a state-action pair $(s,a)$. Applying equation (18) in the proof of Lemma (9) (Jin et al., 2020) with its failure-probability parameter to $K\delta/(|\mathcal{S}||\mathcal{A}|)$, we obtain, with probability at least $1 - \delta/(|\mathcal{S}||\mathcal{A}|)$,

$$\sum_{t=1}^{T}\left(\hat{\ell}_t(s,a) - \frac{q_t(s,a)}{u_t(s,a)}\ell_t(s,a)\right) \leq \frac{1}{2\gamma}\ln\frac{|\mathcal{S}||\mathcal{A}|}{\delta},$$

for each state-action pair $(s,a)$, after taking a union bound over the whole state-action space.

Given that $q_t(s,a) \leq u_t(s,a)$ with probability at least $1 - \frac{\delta}{2}$, and $\sum_{s,a}q(s,a) = K$, we thus have

$$\begin{aligned}
\sum_{t=1}^{T}\left\langle q, \hat{\ell}_t - \ell_t\right\rangle &\leq \sum_{t,s,a}q(s,a)\ell_t(s,a)\left(\frac{q_t(s,a)}{u_t(s,a)} - 1\right) + \sum_{s,a}\frac{q(s,a)\ln\frac{|\mathcal{S}||\mathcal{A}|}{\delta}}{2\gamma} \\
&\leq \frac{K}{2\gamma}\ln\frac{|\mathcal{S}||\mathcal{A}|}{\delta},
\end{aligned}$$

with probability at least $1 - \frac{3\delta}{2}$. Taking $q = q^*$ finishes the proof.

□

### C.3 Proof of Theorem 2

Theorem 7 shows the high-probability upper bound of our TP-FTRL algorithm for the likelihood-calibrated confidence sets when $\kappa = 0$, which completes the proof of Theorem 2

**Theorem 7.** *Suppose $\kappa = 0$. Under the event of Lemma 2, with probability at least $1 - \frac{11\delta}{2}$, under adversarial bandit feedback loss functions and unknown transitions, TP-FTRL with $\eta = \gamma = \sqrt{\frac{2K \ln\left(\frac{|A|^2|\mathcal{S}|^3}{\delta}\right)}{3|A||\mathcal{S}|T}}$ ensures*

$$\begin{aligned} Reg_T(\pi^*) &= \sum_{t=1}^{T} \langle q^{P,\pi_t} - q^{P,\pi^*}, \ell_t \rangle \\ &\leq \mathcal{O}\Bigg( K^{\frac{3}{2}} \sqrt{|\mathcal{S}||\mathcal{A}|T \ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})} + \frac{K \ln(|A||\mathcal{S}|^2)}{\eta} + \frac{\eta}{2}|\mathcal{S}||\mathcal{A}|T + \frac{\eta K}{4\gamma} \ln \frac{K}{\delta} \\ &\qquad + \gamma|\mathcal{S}||\mathcal{A}|T + \frac{K}{\gamma} \ln \frac{|\mathcal{S}||\mathcal{A}|}{\delta} \Bigg) \\ &= \tilde{\mathcal{O}}(K^{\frac{3}{2}} \sqrt{|\mathcal{S}||\mathcal{A}|T}). \end{aligned}$$

*Proof.* Lemma 17 shows the bound of Error reg 1, while Lemma 18 bounds Bias 1. Bias 2 and Error reg 1 are standard and bounded by Lemma 12 and Lemma 8, respectively. We then finish this proof. □

**Lemma 13.** *Suppose $\kappa = 0$. Under the event of Lemma 2, for all episodes $t \in [T]$, all $\hat{P} \in \widehat{\mathcal{C}}_t$, and $(s,a) \in \mathcal{S}_k \times \mathcal{A}$, we have for all $k = 0, \ldots, K-1$,*

$$\sum_{s'} |\hat{P}(s'|s,a) - P(s'|s,a)| = \mathcal{O}\left( \sqrt{\frac{\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s,a)}} \right).$$

*Proof.* For any $\hat{P} \in \widehat{\mathcal{C}}_t$, and for any $(s,a) \in \mathcal{S}_k \times \mathcal{A}$ for all $k = 0, \ldots, K-1$, and $t \in [T]$, we have

$$\mathrm{KL}\left( \bar{P}_t(\cdot|s,a) \,\middle\|\, \hat{P}(\cdot|s,a) \right) \leq \frac{\ln(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s,a)}.$$

Applying Pinsker's inequality, we have

$$\sum_{s'} |\hat{P}(s'|s,a) - \bar{P}(s'|s,a)| \leq \sqrt{2\mathrm{KL}\left( \bar{P}_t(\cdot|s,a) \,\middle\|\, \hat{P}(\cdot|s,a) \right)} \leq \sqrt{\frac{2\ln(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s,a)}}.$$

Since $P \in \widehat{\mathcal{C}}_t$, we also have

$$\sum_{s'} |\bar{P}(s'|s,a) - P(s'|s,a)| \leq \sqrt{\frac{2\ln(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s,a)}}.$$

Thus, combining the above two inequalities gives

$$\sum_{s'} |\hat{P}(s'|s,a) - P(s'|s,a)| \leq \sqrt{\frac{8\ln(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s,a)}}$$

which proves this lemma. □

**Lemma 14.** *Suppose $\kappa = 0$. Under the event of Lemma 2, for all episodes $t \in [T]$, all $\hat{P} \in \widehat{\mathcal{C}}_t$, and $(s,a) \in \mathcal{S}_k \times \mathcal{A}$, we have for all $k = 0, \ldots, K-1$,*

$$|\hat{P}(s'|s,a) - P(s'|s,a)| = \mathcal{O}\left( \sqrt{\frac{P(s'|s,a)\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s,a)}} + \frac{\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s,a)} \right).$$

*Proof.* Applying Hellinger's inequality, we have

$$\sum_{s'}(\sqrt{\hat{P}(s'|s,a)}-\sqrt{\bar{P}(s'|s,a)})^2 \le \mathrm{KL}\left(\bar{P}_t(\cdot|s,a)\,\middle\|\,\hat{P}(\cdot|s,a)\right) \le \frac{\ln(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s,a)},$$

which gives

$$|\sqrt{\hat{P}(s'|s,a)}-\sqrt{\bar{P}(s'|s,a)}| \le \sqrt{\frac{\ln(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s,a)}}.$$

We also have

$$\begin{aligned}|\hat{P}(s'|s,a)-\bar{P}(s'|s,a)| &= (\sqrt{\hat{P}(s'|s,a)}-\sqrt{\bar{P}(s'|s,a)})^2 + \\ &\qquad 2\sqrt{\bar{P}(s'|s,a)}\left|\sqrt{\hat{P}(s'|s,a)}-\sqrt{\bar{P}(s'|s,a)}\right| \\ &\le 2\sqrt{\bar{P}(s'|s,a)\frac{\ln(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s,a)}} + \frac{\ln(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s,a)}.\end{aligned}$$

By the proof of Lemma 7, we finish the proof. □

**Lemma 15.** *Let $n \ge 0$ be an integer and $a_0,\dots,a_n,b_0,\dots,b_n \in \mathbb{R}$. Then*

$$\prod_{i=0}^{n} a_i - \prod_{i=0}^{n} b_i = \sum_{k=0}^{n}\left(\prod_{i=0}^{k-1} a_i\right)(a_k-b_k)\left(\prod_{i=k+1}^{n} b_i\right),$$

*where empty products are defined as* 1.

*Proof.* For $k = 0,\dots,n+1$, define

$$C_k = \left(\prod_{i=0}^{k-1} a_i\right)\left(\prod_{i=k}^{n} b_i\right).$$

For each $k = 0,\dots,n$,

$$C_{k+1} - C_k = \left(\prod_{i=0}^{k-1} a_i\right)(a_k-b_k)\left(\prod_{i=k+1}^{n} b_i\right).$$

Summing these equalities yields

$$\sum_{k=0}^{n}(C_{k+1}-C_k) = C_{n+1}-C_0 = \prod_{i=0}^{n} a_i - \prod_{i=0}^{n} b_i,$$

which proves the claim. □

**Lemma 16.** *Suppose $\kappa = 0$. Under the event of Lemma 2, with probability at least $1-\frac{5\delta}{2}$, the following holds for every collection of transition functions $\{P_t^s\}_{s\in\mathcal{S}}$ satisfying $P_t^s \in \widehat{\mathcal{C}}_t$ for all $s$:*

$$\sum_{t\in[T]}\sum_{s,a}|q^{P_t^s,\pi_t}(s,a)-q_t(s,a)| = \mathcal{O}\left(K^{\frac{3}{2}}\sqrt{|\mathcal{S}||\mathcal{A}|T\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}\right).$$

*Proof.* To simplify notation, we let $q_t^s = q^{P_t^s,\pi_t}$. By Equation (28), for any state-action pair $(s,a)$, we have

$$|q_t(s,a)-q_t^s(s,a)| = \pi_t(a|s)\sum_{\{(s_j,a_j)\}_{j=0}^{k-1}}\prod_{i=0}^{k-1}\pi_t(a_i|s_i)|\prod_{i=0}^{k-1}P(s_{i+1}|s_i,a_i) - \prod_{i=0}^{k-1}P_t^s(s_{i+1}|s_i,a_i)|$$

Let $e_t(s'|s,a) = |P_t^s(s'|s,a) - P(s'|s,a)|$. By Lemma 15, we have

$$
\begin{aligned}
&|\prod_{i=0}^{k-1} P(s_{i+1}|s_i,a_i) - \prod_{i=0}^{k-1} P_t^s(s_{i+1}|s_i,a_i)| \\
&= \sum_{j=0}^{k-1} e_t(s_{j+1}|s_j,a_j) \prod_{i=0}^{j-1} P(s_{i+1}|s_i,a_i) \prod_{i=j+1}^{k-1} P_t^s(s_{i+1}|s_i,a_i).
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
&|q_t^s(s,a) - q_t(s,a)| \\
&\le \pi_t(a|s) \sum_{\{(s_m,a_m)\}_{m=0}^{k-1}} \prod_{i=0}^{k-1} \pi_t(a_i|s_i) \sum_{j=0}^{k-1} e_t(s_{j+1}|s_j,a_j) \prod_{i=0}^{j-1} P(s_{i+1}|s_i,a_i) \prod_{i=j+1}^{k-1} P_t^s(s_{i+1}|s_i,a_i) \\
&= \sum_{j=0}^{k-1} \sum_{(s_j,a_j,s_{j+1})} e_t(s_{j+1}|s_j,a_j) \left( \pi_t(a_j|s_j) \sum_{\{(s_m,a_m)\}_{m=0}^{j-1}} \prod_{i=0}^{j-1} \pi_t(a_i|s_i) P(s_{i+1}|s_i,a_i) \right) \\
&\qquad \cdot \left( \pi_t(a|s) \sum_{\{(s_m,a_m)\}_{m=j+1}^{k-1}} \prod_{i=j+1}^{k-1} \pi_t(a_i|s_i) P_t^s(s_{i+1}|s_i,a_i) \right) \\
&= \sum_{j=0}^{k-1} \sum_{(s_j,a_j,s_{j+1})} e_t(s_{j+1}|s_j,a_j) q_t(s_j,a_j) q_t^s(s,a|s_{j+1}). \qquad (30)
\end{aligned}
$$

Note that $q_t^s(s,a|s_{j+1}) \le q_t(s,a|s_{j+1}) + |q_t^s(s,a|s_{j+1}) - q_t(s,a|s_{j+1})|$. Following the reasoning of Equation (30) and viewing stage $j+1$ as a nominal first stage, we have

$$
\begin{aligned}
|q_t^s(s,a|s_{j+1}) - q_t(s,a|s_{j+1})| &\le \sum_{i=j+1}^{k-1} \sum_{(s_i,a_i,s_{i+1})} e_t(s'_{i+1}|s'_i,a'_i) q_t(s'_i,a'_i|s_{j+1]}) q_t^s(s,a|s'_{i+1}) \\
&\le \pi_t(a|s) \sum_{i=j+1}^{k-1} \sum_{(s'_i,a'_i,s'_{i+1})} e_t(s'_{i+1}|s'_i,a'_i) q_t(s'_i,a'_i|s_{j+1}) \qquad (31)
\end{aligned}
$$

Hence, we have

$$
\begin{aligned}
&\sum_{t\in[T]}\sum_{k=0}^{K-1}\sum_{(s,a)\in\mathcal{S}_k\times\mathcal{A}}|q_t^s(s,a)-q_t(s,a)| \\
&\le \sum_{t\in[T]}\sum_{k=0}^{K-1}\sum_{j=0}^{k-1}\sum_{(s_j,a_j,s_{j+1})} e_t(s_{j+1}|s_j,a_j)q_t(s_j,a_j)\left(\sum_{(s,a)\in\mathcal{S}_k\times\mathcal{A}} q_t(s,a|s_{j+1})\right) \\
&+ \sum_{t\in[T]}\sum_{k=0}^{K-1}\sum_{j=0}^{k-1}\sum_{(s_j,a_j,s_{j+1})} e_t(s_{j+1}|s_j,a_j)q_t(s_j,a_j)\sum_{i=j+1}^{k-1}\sum_{(s'_i,a'_i,s'_{i+1})} e_t(s'_{i+1}|s'_i,a'_i)q_t(s'_i,a'_i|s_{j+1}) \\
&\qquad\cdot\left(\sum_{(s,a)\in\mathcal{S}_k\times\mathcal{A}}\pi_t(a|s)\right) \\
&\le \underbrace{\sum_{t\in[T]}\sum_{k=0}^{K-1}\sum_{j=0}^{k-1}\sum_{(s_j,a_j,s_{j+1})} e_t(s_{j+1}|s_j,a_j)q_t(s_j,a_j)}_{\textbf{Term 1}} + \underbrace{\sum_{t\in[T]}\sum_{k=0}^{K-1}\sum_{j=0}^{k-1}|\mathcal{S}_k|}_{} \\
&\underbrace{\cdot\sum_{s_j,a_j,s_{j+1}} e_t(s_{j+1}\mid s_j,a_j)\,q_t(s_j,a_j)\sum_{i=j+1}^{k-1}\sum_{s'_i,a'_i,s'_{i+1}} e_t(s'_{i+1}\mid s'_i,a'_i)\,q_t(s'_i,a'_i\mid s_{j+1})}_{\textbf{Term 2}}
\end{aligned}
$$

By Lemma 5 and Lemma 13, for **Term 1**, we have

$$
\begin{aligned}
\textbf{Term 1} &= \mathcal{O}\left(\sum_{t\in[T]}\sum_{k=0}^{K-1}\sum_{j=0}^{k-1}\sum_{s_j,a_j} q_t(s_j,a_j)\sqrt{\frac{\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s_j,a_j)}}\right) \\
&\le \mathcal{O}\left(\sum_{k=0}^{K-1}\sum_{j=0}^{k-1}\sqrt{|\mathcal{S}_k||A|T}\sqrt{\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}\right) \\
&\le \mathcal{O}\left(\sum_{k=0}^{K-1} k\sqrt{|\mathcal{S}_k||A|T}\sqrt{\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}\right). \qquad (32)
\end{aligned}
$$

Applying the Cauchy–Schwarz inequality to $\sum_{k=0}^{K-1} k\sqrt{|\mathcal{S}_k|}$, we have

$$
\begin{aligned}
\sum_{k=0}^{K-1} k\sqrt{|\mathcal{S}_k|} &\le \sqrt{\left(\sum_{k=0}^{K-1}k^2\right)\left(\sum_{k=0}^{K-1}|S_k|\right)} \\
&\le \sqrt{\frac{K(K-1)(2K-1)}{6}|\mathcal{S}|} \le K^{\frac{3}{2}}\sqrt{|\mathcal{S}|}.
\end{aligned}
$$

Thus, Equation (32) becomes $\mathcal{O}\left(K^{\frac{3}{2}}\sqrt{|\mathcal{S}||\mathcal{A}|T\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}\right)$.

Now we bound **Term 2**. By Lemma 14, we define:

$$
\epsilon_j = \underbrace{\sqrt{\frac{P(s_{j+1}|s_j,a_j)\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s_j,a_j)}}}_{\epsilon_j^1} + \underbrace{\frac{\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s_j,a_j)}}_{\epsilon_j^2}.
$$

For $\epsilon_j$, we use trivial bounds $\epsilon_j \le 1$. We bound $\epsilon_j\epsilon_i$ as follows:

$$
\epsilon_j\epsilon_i = (\epsilon_j^1+\epsilon_j^2)\epsilon_i \le \epsilon_j^1\epsilon_i + \epsilon_j^2 = \epsilon_j^1\epsilon_i^1 + \epsilon_j^1\epsilon_i^2 + \epsilon_j^2
$$

For $\epsilon_j^1\epsilon_i^2 = \sqrt{\frac{P(s_{j+1}|s_j,a_j)\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s_j,a_j)}}\frac{\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s_i,a_i)}$, we have

$$\epsilon_j^1\epsilon_i^2 \le \frac{\frac{\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s_j,a_j)} + P(s_{j+1}|s_j,a_j)(\frac{\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s_i,a_i)})^2}{2} \le \frac{\frac{\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s_j,a_j)} + P(s_{j+1}|s_j,a_j)\frac{\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s_i,a_i)}}{2}.$$

Since $\sum_{s_j,a_j,s_{j+1}} q_t(s_j,a_j)P(s_{j+1}|s_j,a_j)q_t(s_i',a_i'|s_{j+1}) = q_t(s_i',a_i')$, we have

**Term 2**

$$\begin{aligned}
&\le \sum_{t=1}^{T}\sum_{k=0}^{K-1}|\mathcal{S}_k|\sum_{j=0}^{k-1}\sum_{i=j+1}^{k-1}\sum_{s_i',a_i',s_{i+1}'} q_t(s_j,a_j)q_t(s_i',a_i'|s_{j+1})\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta}) \\
&\quad\cdot\sqrt{\frac{P(s_{j+1}|s_j,a_j)}{N_t(s_j,a_j)}}\sqrt{\frac{P(s_{i+1}'|s_i',a_i')}{N_t(s_i',a_i')}} + \sum_{t=1}^{T}\sum_{k=0}^{K-1}|\mathcal{S}_k|\sum_{j=0}^{k-1}\sum_{i=j+1}^{k-1}\sum_{s_j,a_j,s_{j+1}}\sum_{s_i',a_i',s_{i+1}'} q_t(s_i',a_i') \\
&\quad\cdot\frac{\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s_i',a_i')} + \sum_{t=1}^{T}\sum_{k=0}^{K-1}|\mathcal{S}_k|\sum_{j=0}^{k-1}\sum_{i=j+1}^{k-1}\sum_{s_j,a_j,s_{j+1}}\sum_{s_i',a_i',s_{i+1}'} q_t(s_j,a_j)\frac{\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s_j,a_j)}
\end{aligned}$$

By Lemma 5, the last two terms are both of order $\mathcal{O}(\ln T)$. The first term is $\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})$ multiplied by

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{k=0}^{K-1}|\mathcal{S}_k|\sum_{j=0}^{k-1}\sum_{i=j+1}^{k-1}\sum_{\substack{s_j,a_j,s_{j+1}\\ s_i',a_i',s_{i+1}'}} q_t(s_j,a_j)q_t(s_i',a_i'|s_{j+1})\sqrt{\frac{P(s_{j+1}|s_j,a_j)}{N_t(s_j,a_j)}}\sqrt{\frac{P(s_{i+1}'|s_i',a_i')}{N_t(s_i',a_i')}} \\
&\le \mathcal{O}\Bigg(\sum_{k=0}^{K-1}\sum_{j=0}^{k-1}\sum_{i=j+1}^{k-1}|\mathcal{S}_k|\sqrt{\sum_{\substack{t,s_j,a_j,s_{j+1}\\ s_i',a_i',s_{i+1}'}}\frac{q_t(s_j,a_j)P(s_{i+1}'|s_i',a_i')q_t(s_i',a_i'|s_{j+1})}{N_t(s_j,a_j)}} \qquad (33) \\
&\qquad\cdot\sqrt{\sum_{\substack{t,s_j,a_j,s_{j+1}\\ s_i',a_i',s_{i+1}'}}\frac{q_t(s_j,a_j)P(s_{j+1}|s_j,a_j)q_t(s_i',a_i'|s_{j+1})}{N_t(s_i',a_i')}}\Bigg) \\
&= \mathcal{O}\Bigg(\sum_{k=0}^{K-1}\sum_{j=0}^{k-1}\sum_{i=j+1}^{k-1}|\mathcal{S}_k|\sqrt{\sum_{t,s_j,a_j,s_{j+1}}\frac{q_t(s_j,a_j)}{N_t(s_j,a_j)}}\sqrt{\sum_{t,s_i',a_i',s_{i+1}'}\frac{q_t(s_i',a_i')}{N_t(s_i',a_i')}}\Bigg) \\
&\le \mathcal{O}\left(\sum_{k=0}^{K-1}\sum_{j=0}^{k-1}\sum_{i=j+1}^{k-1}|\mathcal{S}_k|\sqrt{|\mathcal{S}_j||\mathcal{S}_{j+1}||A|\ln T}\sqrt{|\mathcal{S}_i||\mathcal{S}_{i+1}||A|\ln T}\right) \qquad (34) \\
&\le \mathcal{O}\left(|\mathcal{S}|^3|A|\ln T\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})\right)
\end{aligned}$$

We use the Cauchy–Schwarz inequality for Equations (33) and (34). Thus, ERROR REG 1 $= \mathcal{O}(K^{\frac{3}{2}}\sqrt{|\mathcal{S}||\mathcal{A}|T\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})})$. Conditioning on Lemma 5 and Lemma 13, the events happen with probability at least $1-\frac{5\delta}{2}$, which finishes the proof.

□

**Lemma 17.** *Suppose $\kappa = 0$. Under the event of Lemma 2, with probability at least $1-\frac{5\delta}{2}$, TP-FTRL ensures* ERROR REG $1 = \mathcal{O}\left(K^{\frac{3}{2}}\sqrt{|\mathcal{S}||\mathcal{A}|T\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}\right)$.

*Proof.* Since $\hat{q}_t \in \widehat{C}_t$, by Lemma 16, we have

$$\begin{aligned}\textsc{Error reg} \ 1 &= \sum_{t=1}^{T} \langle q_t - \widehat{q}_t, \ell_t \rangle \\ &\leq \sum_{t=1}^{T} \sum_{k=0}^{K-1} \sum_{(s,a)\in\mathcal{S}_k\times\mathcal{A}} |q_t(s,a) - \hat{q}_t(s,a)| \\ &= \mathcal{O}\left( K^{\frac{3}{2}} \sqrt{|\mathcal{S}||\mathcal{A}|T \ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})} \right),\end{aligned}$$

with probability at least $1 - \frac{5\delta}{2}$. □

**Lemma 18.** *Suppose $\kappa = 0$. Under the event of Lemma 2, with probability at least $1 - \frac{7\delta}{2}$, TP-FTRL ensures* $\textsc{Bias} \ 1 = \mathcal{O}\left( K^{\frac{3}{2}} \sqrt{|\mathcal{S}||\mathcal{A}|T \ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})} + \gamma|\mathcal{S}||\mathcal{A}|T \right)$.

*Proof.* By the proof of Lemma 11, we have

$$\textsc{Bias} \ 1 \leq \sum_{t=1}^{T} \sum_{s,a} |(u_t(s,a) - q_t(s,a)| + \gamma|\mathcal{S}||\mathcal{A}|T,$$

with probability at least $1 - \delta$. Then by Lemma 17, we have the final bound as follows:

$$\textsc{Bias} \ 1 = \mathcal{O}\left( K^{\frac{3}{2}} \sqrt{|\mathcal{S}||\mathcal{A}|T \ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})} + \gamma|\mathcal{S}||\mathcal{A}|T \right),$$

with probability at least $1 - \frac{7\delta}{2}$. □

### C.4 Proof of Theorem 3

Theorem 8 shows the high-probability upper bound of our TP-FTRL algorithm for the likelihood-calibrated confidence sets, which completes the proof of Theorem 3

**Theorem 8.** *Under the event of Lemma 2, with probability at least $1 - \frac{11\delta}{2}$, under adversarial bandit feedback loss functions and unknown transitions, TP-FTRL with $\eta = \gamma = \sqrt{\frac{2K \ln\left(\frac{|A|^2|\mathcal{S}|^3}{\delta}\right)}{3|A||\mathcal{S}|T}}$ ensures*

$$\begin{aligned}Reg_T(\pi^*) &= \sum_{t=1}^{T} \langle q^{P,\pi_t} - q^{P,\pi^*}, \ell_t \rangle \\ &\leq \mathcal{O}\Bigg( K^{\frac{3}{2}} \sqrt{(1+\kappa)|\mathcal{S}||\mathcal{A}|T \ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})} + \frac{K \ln(|A||\mathcal{S}|^2)}{\eta} + \frac{\eta}{2}|\mathcal{S}||\mathcal{A}|T + \frac{\eta K}{4\gamma} \ln \frac{K}{\delta} \\ &\qquad + \gamma|\mathcal{S}||\mathcal{A}|T + \frac{K}{\gamma} \ln \frac{|\mathcal{S}||\mathcal{A}|}{\delta} \Bigg) \\ &= \tilde{\mathcal{O}}(K^{\frac{3}{2}} \sqrt{(1+\kappa)|\mathcal{S}||\mathcal{A}|T}).\end{aligned}$$

*Proof.* Lemma 20 shows the bound of Error reg 1, while Lemma 21 bounds Bias 1. Bias 2 and Error reg 1 are standard and bounded by Lemma 12 and Lemma 8, respectively. We then finish this proof. □

**Lemma 19.** *Under the event of Lemma 2, for all episodes $t \in [T]$, all $\hat{P} \in \widehat{\mathcal{C}}_t$, and $(s,a) \in \mathcal{S}_k \times \mathcal{A}$, we have for all $k = 0, \ldots, K-1$,*

$$\sum_{s'} |\hat{P}(s'|s,a) - P(s'|s,a)| = \mathcal{O}\left( \sqrt{(1+\kappa)\frac{\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s,a)}} \right), \tag{35}$$

*and*

$$|\hat{P}(s'|s,a) - P(s'|s,a)| = \mathcal{O}\left(\sqrt{(1+\kappa)\frac{P(s'|s,a)\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s,a)}} + \frac{(1+\kappa)\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s,a)}\right). \quad (36)$$

*Proof.* Following the proofs of Lemma 13 and Lemma 14, we substitute $\frac{\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s,a)}$ with $(1+\kappa)\frac{\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s,a)}$. The results then follow. □

**Lemma 20.** *Under the event of Lemma 2, with probability at least* $1-\frac{5\delta}{2}$*, TP-FTRL ensures* ERROR REG $1 = \mathcal{O}\left(K^{\frac{3}{2}}\sqrt{(1+\kappa)|\mathcal{S}||\mathcal{A}|T\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}\right)$.

*Proof.* By the proof of Lemma 16, with probability at least $1-\frac{5\delta}{2}$, we have the following:

$$\textbf{Term 1} \leq \mathcal{O}\left(K^{\frac{3}{2}}\sqrt{(1+\kappa)|\mathcal{S}||\mathcal{A}|T\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}\right),$$

$$\textbf{Term 2} \leq \mathcal{O}\left((1+\kappa)|\mathcal{S}|^3|A|\ln T\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})\right).$$

Thus, ERROR REG $1 = \mathcal{O}\left(K^{\frac{3}{2}}\sqrt{(1+\kappa)|\mathcal{S}||\mathcal{A}|T\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}\right)$.

□

**Lemma 21.** *Under the event of Lemma 2, with probability at least* $1-\frac{7\delta}{2}$*, TP-FTRL ensures* BIAS $1 = \mathcal{O}\left(K^{\frac{3}{2}}\sqrt{(1+\kappa)|\mathcal{S}||\mathcal{A}|T\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})} + \gamma|\mathcal{S}||\mathcal{A}|T\right)$.

*Proof.* By the proof of Lemma 11, we have

$$\text{BIAS } 1 \leq \sum_{t=1}^{T}\sum_{s,a}|(u_t(s,a) - q_t(s,a)| + \gamma|\mathcal{S}||\mathcal{A}|T,$$

with probability at least $1-\delta$. Then by the proof of Lemma 20, we have the final bound as follows:

$$\text{BIAS } 1 = \mathcal{O}\left(K^{\frac{3}{2}}\sqrt{(1+\kappa)|\mathcal{S}||\mathcal{A}|T\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})} + \gamma|\mathcal{S}||\mathcal{A}|T\right),$$

with probability at least $1-\frac{7\delta}{2}$. □

### C.5 PROOF OF THEOREM 4

Theorem 9 shows the high-probability upper bound of our TP-FTRL algorithm for the nested confidence sets $\mathcal{C}$, which combine the empirical and likelihood-calibrated confidence sets ($\bar{\mathcal{C}}$ and $\widehat{\mathcal{C}}$). This completes the proof of Theorem 4

**Theorem 9.** *Under the event of Proposition 2, with probability at least* $1-6\delta$*, under adversarial bandit feedback loss functions and unknown transitions, TP-FTRL with* $\eta = \gamma = \sqrt{\frac{2K\ln\left(\frac{|A|^2|\mathcal{S}|^3}{\delta}\right)}{3|A||\mathcal{S}|T}}$

*ensures*

$$
\begin{aligned}
Reg_T(\pi^*) &= \sum_{t=1}^{T} \langle q^{P,\pi_t} - q^{P,\pi^*}, \ell_t \rangle \\
&\leq \mathcal{O}\Bigg( \min\left\{ K^{\frac{3}{2}} \sqrt{(1+\kappa)|\mathcal{S}||\mathcal{A}|T \ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})},\ K|\mathcal{S}|\sqrt{|\mathcal{A}|T\ln(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta})} \right\} + \frac{K\ln(|A||\mathcal{S}|^2)}{\eta} \\
&\qquad + \frac{\eta}{2}|\mathcal{S}||\mathcal{A}|T + \frac{\eta K}{4\gamma}\ln\frac{K}{\delta} + \gamma|\mathcal{S}||\mathcal{A}|T + \frac{K}{\gamma}\ln\frac{|\mathcal{S}||\mathcal{A}|}{\delta} \Bigg) \\
&= \tilde{\mathcal{O}}\left( K\sqrt{|\mathcal{S}||\mathcal{A}|T}\ \min\left\{ \sqrt{K(1+\kappa)},\ \sqrt{|\mathcal{S}|} \right\} \right).
\end{aligned}
$$

*Proof.* By Lemma 22 and the proofs of Lemma 4 and Lemma 20, we have

$$
\textsc{Error reg} \ 1 = O\Bigg( \min\left\{ K^{\frac{3}{2}} \sqrt{(1+\kappa)|\mathcal{S}||\mathcal{A}|T \ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})},\ K|\mathcal{S}|\sqrt{|\mathcal{A}|T\ln(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta})} \right\} \Bigg),
$$

with probability at least $1-3\delta$.

$$
\textsc{Bias} \ 1 = \mathcal{O}\Bigg( \min\left\{ K^{\frac{3}{2}} \sqrt{(1+\kappa)|\mathcal{S}||\mathcal{A}|T \ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})},\ K|\mathcal{S}|\sqrt{|\mathcal{A}|T\ln(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta})} \right\} + \gamma|\mathcal{S}||\mathcal{A}|T \Bigg),
$$

with probability at least $1-4\delta$.

$$
\textsc{Error reg} \ 2 = \mathcal{O}\left( \frac{K\ln(|A||\mathcal{S}|^2)}{\eta} + \frac{\eta}{2}|\mathcal{S}||\mathcal{A}|T + \frac{\eta K}{4\gamma}\ln\frac{K}{\delta} \right),
$$

with probability at least $1-2\delta$.

$$
\textsc{Bias} \ 2 = \mathcal{O}\left( \frac{K}{\gamma}\ln\frac{|\mathcal{S}||\mathcal{A}|}{\delta} \right),
$$

with probability at least $1-2\delta$.

Combining them shows the event happens with probability at least $1-6\delta$, which finishes the proof. □

**Lemma 22.** *Under the event of Proposition 2, with probability at least $1-\delta$, for all episodes $t \in [T]$, all $\hat{P} \in \widehat{\mathcal{C}}_t$, and $(s,a) \in \mathcal{S}_k \times \mathcal{A}$, we have for all $k = 0, \ldots, K-1$,*

$$
\sum_{s'} |\hat{P}(s'|s,a) - P(s'|s,a)| = \mathcal{O}\left( \min\left\{ \sqrt{(1+\kappa)\frac{\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s,a)}},\ \sum_{s'} \bar{\epsilon}_t(s'|s,a) \right\} \right), \tag{37}
$$

*and*

$$
|\hat{P}(s'|s,a) - P(s'|s,a)| = \mathcal{O}\left( \min\{\hat{\epsilon}_t(s'|s,a),\ \bar{\epsilon}_t(s'|s,a)\} \right), \tag{38}
$$

*where* $\hat{\epsilon}_t(s'|s,a) = \sqrt{(1+\kappa)\frac{P(s'|s,a)\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s,a)}} + \frac{(1+\kappa)\ln(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N_t(s,a)}$ *and* $\bar{\epsilon}_t(s'|s,a) = \sqrt{\frac{P(s'|s,a)\ln(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta})}{\max\{1,N_t(s,a)\}}} + \frac{\ln(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta})}{\max\{1,N_t(s,a)\}}$.

*Proof.* By Lemma 7 and Lemma 19, we have Equation 38. Then summing $\bar{\epsilon}_t(s'|s,a)$ over $s'$ proves Equation 37. □