Doubly Regularized Markov Decision Process for Robust Reinforcement Learning

Yiting He*‡ Zhishuai Liu†‡ Pan Xu§

Abstract

Empirical successes show that regularization improves the stability and efficiency of reinforcement learning (RL), with applications in robotics and post-training of large language models. Yet, theoretical analyses of regularized Markov decision processes (MDPs) have mostly been confined to the standard RL setting. In this work, we investigate regularized MDPs through the lens of robust RL. We introduce a doubly regularized MDP framework that combines policy and dynamics regularization to enable robust policy learning against reward and dynamics perturbations. Within this framework, we develop an optimism-based online algorithm and provide the first finite-sample regret guarantees in both tabular and rich-observation settings, where the state-action space may be continuous. Our results show that algorithms for doubly regularized MDPs are as sample-efficient as well-studied robust MDP algorithms, while additionally benefiting from the flexibility of soft policies. Finally, experiments on CartPole demonstrate that our approach efficiently and effectively handles function approximation and exploration in large state-action spaces, achieving robust performances.

1 INTRODUCTION

Regularizations have been widely studied in reinforcement learning (RL) literature. Practically, existing works leveraging policy regularization have demonstrated tremendous successes in real world applications such as game playing (Mnih et al., 2016), robotics control (Schulman et al., 2015, 2017) and post-training of large language models (Bai et al., 2022; Ouyang et al., 2022). Theoretically, extensive studies (Geist et al., 2019; Yang et al., 2019; Vieillard et al., 2020; Zhao et al., 2025) have investigated the policy regularized MDP in the standard RL setting. Moreover, recent works (Derman et al., 2021; Eysenbach and Levine, 2021; Husain et al., 2021) also point out its relation with reward-robust MDPs. Another line of works (Yang et al., 2023; Zhang et al., 2024; He et al., 2025) propose the regularized robust MDP (RRMDP) framework, where a dynamics regularization is added to the objective function to account for robustness against potential dynamics shift. They have shown that the dynamics regularization leads to more efficient estimation procedures of value functions compared with the conventional distributionally robust MDP framework (Iyengar, 2005; Nilim and El Ghaoui, 2005).

To combine the strength of regularization, in this work, we unify the policy and dynamics regularization in a doubly regularized MDP framework, and investigate it through the lens of robust

*University of Science and Technology of China; email: heyiting@mail.ustc.edu.cn
†Duke University; email: zhishuai.liu@duke.edu
‡Equal contribution
§Duke University; email: pan.xu@duke.edu

reinforcement learning. In particular, we first establish the framework in the tabular setting where the state and action spaces are finite. Unlike prior works (Yang et al., 2023; Zhang et al., 2024; Panaganti et al., 2024) assuming access to a fixed dataset or a simulator, we focus on the more realistic online setting, where the agent can only gather data through interaction with the environment. In this setting, the agent faces the unique challenge of exploration, that is, efficiently acquiring information about the environment through strategically selecting the policy. Existing results under this online setting often rely on restrictive structural assumptions, such as the fail-state condition in Lu et al. (2024); Liu and Xu (2024a); Liu et al. (2024), which applies only to uncertainty sets defined by Total Variation distance and does not extend to more general f-divergences. Recently, He et al. (2025) identified the information deficit issue as the central obstacle in robust RL, arising from the need to generalize experience under distribution shifts to perturbed environments while only interacting with the source environment. Building on this, we adopt the bounded visitation measure ratio assumption proposed by He et al. (2025), extending the analysis to handle the double regularization of our framework. Further, to account for large state and action spaces in practical scenarios, we extend our doubly regularized MDP framework to incorporate function approximation for generalization. We summarize our main contribution as follows:

- We establish a novel doubly regularized MDP framework that incorporates both transition regularization and policy regularization, and show that this framework achieves robustness with respect to both transitions and rewards.
- Under this framework, we propose an online learning algorithm Robust Soft Policy Value Iteration (RSPVI) with general f-divergence based transition regularization terms and several policy regularization formulations. We establish the first finite sample regret bounds for doubly regularized MDPs.
- We further extend RSPVI to the setting with linear function approximation. With fewer structural assumptions than prior work, we provide the first finite-sample regret guarantees applicable to broader settings defined by general f-divergences.
- In stark contrast to the conventional robust MDP framework, the optimal policies under doubly regularized MDPs are stochastic. We empirically evaluate RSPVI in the CartPole environment, showing that the stochastic policies learned by RSPVI facilitate exploration and posses robustness performances against environment perturbation.

Notations For $H \in \mathbb{Z}_+$, we denote $[H] = \{1, \cdots, H\}$. For any set $\mathcal{S}$, define $\Delta(\mathcal{S})$ as the set of probability distributions over $\mathcal{S}$. For a convex function $f : [0, +\infty) \to (-\infty, +\infty]$ with $f(x)$ finite for $x > 0$, $f(1) = 0$ and $f(0) = \lim_{t\to 0^+} f(t)$. For $P, Q \in \Delta(\mathcal{S})$ and $P \ll Q$, the f-divergence of P from Q is defined as $D_f(P\|Q) = \int_\Omega f(\frac{\mathrm{d}P}{\mathrm{d}Q})\,\mathrm{d}Q$. In this paper, we focus on total variation (TV) distance with $f(t) = \frac{1}{2}|t-1|$, Kullback-Leibler (KL) divergence with $f(t) = t\ln t$, and χ^2-divergence with $f(t) = (t-1)^2$. We use $\mathcal{O}(\cdot)$ to hide absolute constant factors and $\widetilde{\mathcal{O}}(\cdot)$ to further hide logarithmic factors. We denote $\mathrm{proj}_C(x) = \arg\inf_{c\in C} \|c - x\|^2$ as the projection of x onto the set C. We denote the support function of a set $\mathcal{Z} \subseteq \mathbb{R}^n$ as $\sigma_{\mathcal{Z}}(\boldsymbol{y}) = \max_{\boldsymbol{a}\in\mathcal{Z}} \langle \boldsymbol{a}, \boldsymbol{y}\rangle$.

2 RELATED WORK

Robust MDPs Robust MDPs were first introduced as a control problem with a known nominal transition model (Iyengar, 2005; Nilim and El Ghaoui, 2005; Xu and Mannor, 2006; Wiesemann et al., 2013; Mannor et al., 2016), where robust policies are derived by solving a constrained max-min optimization problem. Subsequent work extended robust MDPs to the learning setting, both with

access to a generative model allowing the agent to sample transitions for arbitrary state-action pair (Zhou et al., 2021; Yang et al., 2022; Panaganti and Kalathil, 2022; Shi et al., 2024) and with pre-collected datasets that provide adequate coverage of the optimal policy under the perturbed environment (Shi and Chi, 2024; Panaganti et al., 2022; Blanchet et al., 2023; Wang et al., 2024a; Liu and Xu, 2024b, 2025). More recently, Lu et al. (2024); Liu and Xu (2024a) studied robust MDPs in the online setting under the fail-state or vanishing minimal-value structure assumption. He et al. (2025) identified the key challenge of online robust learning as the information deficit issue, and addressed this by introducing a maximum visitation ratio assumption, extending the analysis to the general f-divergence setting. Ghosh et al. (2025) employed a variance-aware proof technique to refine the results of He et al. (2025).

Regularized Robust MDPs Unlike the above approaches formulating robust MDPs as a constrained optimization problem, regularized robust MDPs were first introduced by Yang et al. (2023) and Zhang et al. (2024) as an efficient alternative. Yang et al. (2023) focused on the generative model setting, providing a sample complexity analysis and proving the statistical equivalence of the two formulations. Zhang et al. (2024) established the equivalence with risk-sensitive MDPs, analyzed the convergence rate of policy gradient methods, and proposed a value iteration algorithm with general function approximation for the offline setting. Subsequently, Panaganti et al. (2024) extended the framework to general f-divergences with function approximation in the offline setting, and to the Total Variation distance under a fail-state assumption in the hybrid offline-online setting. More recently, Tang et al. (2025) studied linear function approximation for general f-divergences using offline datasets.

Reward Robust MDPs Most existing works focus solely on transition robustness. For studies addressing reward robustness, notable examples include Zhou et al. (2021), Wang et al. (2023), and Wang et al. (2024b). All of these studies formulate reward robustness through an uncertainty set. Specifically, Zhou et al. (2021) proposed a model-based value iteration algorithm using a pre-collected dataset, while Wang et al. (2023) introduced a model-free Q-learning algorithm under the assumption of access to a simulator. Building on this, Wang et al. (2024b) incorporated a variance-reduction technique into Q-learning, also assuming access to a generative model. However, none of these works address the more practical online setting. The work most closely related to ours is Derman et al. (2021), which demonstrates the equivalence between policy regularization and reward robustness. However, it relies on the assumption that the nominal transition is known, which is a fairly strong requirement.

3 PRELIMINARIES

3.1 Problem Formulations

Markov Decision Process (MDP) A finite-horizon Markov Decision Process is denoted as the tuple $(\mathcal{S}, \mathcal{A}, P, r, H)$. Here, $\mathcal{S}$ and $\mathcal{A}$ represent the state and action spaces. We also denote $|\mathcal{A}|$ as the volume of $\mathcal{A}$, where $|\mathcal{A}|$ is the number of actions for discrete action space and $|\mathcal{A}| = \int_{\mathcal{A}} 1 \, da$ for continuous space. The transition dynamics are given by $P = \{P_h\}_{h=1}^H$, where each $P_h(\cdot|s,a) : \mathcal{S} \to \Delta(\mathcal{S})$ specifies the probability distribution over the next state when action a is taken at state s and step h. The reward functions are $r = \{r_h\}_{h=1}^H$, with each $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ assumed to be

deterministic. We define the value function and Q-function as

$$V_h^\pi(s) = \mathbb{E}_{\pi,P}\Big[\sum_{t=h}^{H} r_t(s_t,a_t) \,\Big|\, s_h = s\Big],$$
$$Q_h^\pi(s,a) = \mathbb{E}_{\pi,P}\Big[\sum_{t=h}^{H} r_t(s_t,a_t) \,\Big|\, s_h = s, a_h = a\Big].$$

Doubly Regularized MDPs We now introduce the doubly regularized framework. A finite-horizon doubly regularized MDP can be denoted as the tuple $(\mathcal{S}, \mathcal{A}, P^o, r^o, \beta, \mathrm{D}, \Omega^\eta, H)$. Here, D is a probability divergence that quantifies the distribution shift (later we will instantiate D as TV, KL or χ^2-divergence), and β is the corresponding regularization parameter. $\Omega_s^\eta(\pi) \triangleq \Omega^\eta(s, \pi(\cdot|s))$ denotes a penalty function on the chosen policy π at state s, with the η in Ω^η controlling the extent of this penalty. We define the doubly regularized value function and Q-function as

$$V_h^{\pi,\beta,\eta}(s) = \inf_{P_t\in\Delta(\mathcal{S})} \mathbb{E}_{\pi,\{P_t\}_{t=h}^H}\Big[\sum_{t=h}^{H}\Big(r_t(s_t,a_t) + \beta\cdot \mathrm{D}(P_t(\cdot|s_t,a_t)\|P_t^o(\cdot|s_t,a_t)) - \Omega_{s_t}^\eta(\pi_t)\Big) \,\Big|\, s_h = s\Big],$$
$$Q_h^{\pi,\beta,\eta}(s,a) = \inf_{P_t\in\Delta(\mathcal{S})} \mathbb{E}_{\pi,\{P_t\}_{t=h}^H}\Big[\sum_{t=h}^{H}\Big(r_t(s_t,a_t) + \beta\cdot \mathrm{D}(P_t(\cdot|s_t,a_t)\|P_t^o(\cdot|s_t,a_t))\Big) - \sum_{t=h+1}^{H}\Omega_{s_t}^\eta(\pi_t) \,\Big|\, s_h = s, a_h = a\Big].$$

We then define the optimal value function $V_h^{*,\beta,\eta}(s) = \sup_{\pi\in\Pi} V_h^{\pi,\beta,\eta}(s)$ and the optimal Q-function $Q_h^{*,\beta,\eta}(s,a) = \sup_{\pi\in\Pi} Q_h^{\pi,\beta,\eta}(s,a)$ for all $(h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$, where Π is the set of all possible policies. Correspondingly, the optimal policy $\pi^* = \{\pi_h^*\}_{h=1}^H$ is defined as the policy that achieves the optimal value function for all $(h,s)\in[H]\times\mathcal{S}$, that is, $\pi_h^* = \arg\sup_{\pi\in\Pi} V_h^{\pi,\beta,\eta}(s)$.

Dynamic Programming Principles Under the doubly regularized framework, we establish the dynamic programming principle as follows

Proposition 3.1. For doubly regularized tabular MDPs, it holds that for any policy π and any $(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}$,

$$Q_h^{\pi,\beta,\eta}(s,a) = r_h(s,a) + \inf_{P_h\in\Delta(\mathcal{S})}\big\{\mathbb{E}_{s'\sim P_h(\cdot|s,a)}[V_{h+1}^{\pi,\beta,\eta}(s')] + \beta\cdot\mathrm{D}\big(P_h(\cdot|s,a)\big\|P_h^o(\cdot|s,a)\big)\big\}, \tag{3.1}$$
$$V_h^{\pi,\beta,\eta}(s) = -\Omega_s^\eta(\pi_h) + \mathbb{E}_{a\sim\pi_h(\cdot|s)}\big[Q_h^{\pi,\beta,\eta}(s,a)\big]. \tag{3.2}$$

Double Robustness of the Proposed Framework The effectiveness of incorporating regularization on transition deviations and accounting for the worst case in transition robustness has been extensively studied in the literature (Yang et al., 2023; Zhang et al., 2024; Panaganti et al., 2024; He et al., 2025; Tang et al., 2025). Proposition 3.2 further demonstrates that introducing a policy

regularization term is equivalent to considering the worst-case reward within a suitably defined uncertainty set (Derman et al., 2021). Therefore, our proposed doubly regularized framework can achieve double robustness.

Proposition 3.2. When setting the policy regularization term as

$$\Omega_s^\eta(\pi) \triangleq \Omega^\eta(s, \pi(\cdot|s)) = \sigma_{\mathcal{R}_s^\eta(\pi)}(-\pi(\cdot|s))$$

for all $(\pi, s) \in \Pi \times \mathcal{S}$, where $\mathcal{R}^\eta : \mathcal{S} \times \mathcal{A} \times \Pi \to I \subseteq \mathbb{R}$ denotes the uncertainty set within which the reward deviation $\Delta r = \widetilde{r} - r$ may lie, and $\sigma_{\mathcal{R}_s^\eta(\pi)}(-\pi(\cdot|s)) = \sup_{\Delta r \in \mathcal{R}^\eta(\pi)} \langle -\pi(\cdot|s), \Delta r(s, \cdot)\rangle$. Then for any state $s \in \mathcal{S}$ and any function $V : \mathcal{S} \to \mathbb{R}$, we have

$$\begin{aligned}
&\inf_{\substack{\widetilde{r} \in r + \mathcal{R}^\eta(\pi) \\ P \in \Delta(\mathcal{S})}} \mathbb{E}_{a\sim\pi(\cdot|s)} \left[\widetilde{r}(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)}[V(s')] + \beta \cdot \mathrm{D}\big(P(s,a) \big\| P^o(s,a)\big)\right] \\
= &\inf_{P \in \Delta(\mathcal{S})} \mathbb{E}_{a\sim\pi(\cdot|s)} \left[r(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)}[V(s')] + \beta \cdot \mathrm{D}\big(P(s,a) \big\| P^o(s,a)\big) - \Omega_s^\eta(\pi)\right].
\end{aligned}$$

Learning Goal We consider the online setting, where an agent aims to learn the optimal robust policy by interacting with the environment over K episodes. At the start of each episode k, the agent is given an initial state s_1^k, which we assume to be fixed without loss of generality. For the policy π^k selected by the agent based on the history at episode k, we measure its sub-optimality by $V_1^{*,\beta,\eta}(s_1) - V_1^{\pi^k,\beta,\eta}(s_1)$. The cumulative sub-optimality after K episodes, which is commonly referred to as $\mathrm{Regret}(K)$, is defined as

$$\mathrm{Regret}(K) = \sum_{k=1}^{K} \left(V_1^{*,\beta,\eta}(s_1) - V_1^{\pi^k,\beta,\eta}(s_1)\right).$$

4 ALGORITHM

First, we present a meta-algorithm Algorithm 1 for online doubly regularized MDPs, incorporating general f-divergence based regularization terms and various types of policy regularization terms. We then instantiate Algorithm 1 in each case.

Similar to Lu et al. (2024); Liu and Xu (2024a); He et al. (2025), our algorithm Robust Soft Policy Value Iteration (RSPVI) adopts a value iteration framework. We leverage the robust Bellman optimality equation and incorporate the optimism principle (Abbasi-Yadkori et al., 2011) to estimate the robust Q-functions. For the explicit computation expression, we employ the dual formulation corresponding to each f-divergence setting.

4.1 Model Estimation

We use a model based manner to estimate the empirical reward and transition in the tabular setting. In each episode k, after executing policy π^k and collecting a trajectory $\tau^k = (s_1^k, a_1^k, r_1^k, \cdots, s_H^k, a_H^k, r_H^k)$, RSPVI updates the empirical estimations as follows

$$n_h^k(s,a) = \sum_{i=1}^{k} \mathbb{1}\left\{s_h^i = s, a_h^i = a\right\},$$

Algorithm 1 Robust Soft Policy Value Iteration (RSPVI) for tabular setting

Require: transition regularizer $\beta > 0$, policy regularizer Ω^η.

for $k = 1, \cdots, K$ **do**
  $V_{H+1}^{k,\beta,\eta}(\cdot) \leftarrow 0$.
  for $h = H, \cdots, 1$ **do**
    for $\forall\, (s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
      Update Q-function estimation $Q_h^{k,\beta,\eta}(s,a)$ according to (4.2).
    end for
    for $\forall\, s \in \mathcal{S}$ **do**
      Update policy $\pi_h^k(\cdot|s)$ according to (4.3).
      Update value function estimation $V_h^{k,\beta,\eta}(s) = \big\langle Q_h^{k,\beta,\eta}(s,\cdot), \pi_h^k(\cdot|s)\big\rangle - \Omega_s^\eta(\pi_h^k)$.
    end for
  end for
  Collect a trajectory τ^k by executing π^k.
  Update model estimation as (4.1).
end for

$$\widehat{r}_h^{k+1}(s,a) = \frac{\sum_{i=1}^k r_h^i(s,a)\cdot \mathbb{1}\{s_h^i = s, a_h^i = a\}}{n_h^k(s,a) \vee 1}, \tag{4.1}$$
$$\widehat{P}_h^{k+1}(s'|s,a) = \frac{\sum_{i=1}^k \mathbb{1}\{s_h^i = s, a_h^i = a, s_{h+1}^i = s'\}}{n_h^k(s,a) \vee 1}.$$

4.2 Q-function Estimation

The estimation of optimal robust Q-function $Q_h^{k,\beta,\eta}$, $(h,k) \in [H] \times [K]$ is constructed as follows

$$Q_h^{k,\beta,\eta}(s,a) = \min\big\{\mathrm{RB}_h^k(s,a) + b_h^k(s,a), H-h+1\big\}, \tag{4.2}$$

which consists of a robust Bellman estimator $\mathrm{RB}_h^k(s,a)$ and a bonus term $b_h^k(s,a)$. In the following, we instantiate this generic form under different f-divergence settings, providing explicit formulations for robust Bellman estimation as well as bonus design.

Tabular-TV Setting According to the dual formulation for regularized TV setting in Lemma C.3, given the estimated value function $V_{h+1}^{k,\beta,\eta}$ and empirical reward $\widehat{r}_h^k$ and transition $\widehat{P}_h^k$, we choose the robust Bellman estimator $\mathrm{RB}_h^k(s,a)$ and the bonus $b_h^k(s,a)$ as

$$\mathrm{RB}_h^k(s,a) = \widehat{r}_h^k(s,a) - \mathbb{E}_{\widehat{P}_h^k}\Big[\Big(\min_{s\in\mathcal{S}} V_{h+1}^{k,\beta,\eta}(s) + \beta - V_{h+1}^{k,\beta,\eta}(s)\Big)_+\Big](s,a) + \Big(\min_{s\in\mathcal{S}} V_{h+1}^{k,\beta,\eta}(s) + \beta\Big),$$
$$b_h^k(s,a) = 2H\sqrt{\frac{2S\ln(2SAHK/\delta)}{n_h^{k-1}(s,a)\vee 1}}.$$

Tabular-KL Setting According to the dual formulation for regularized KL setting in Lemma C.6, given the estimated value function $V_{h+1}^{k,\beta,\eta}$ and empirical reward $\widehat{r}_h^k$ and transition $\widehat{P}_h^k$, we choose the robust Bellman estimator $\mathrm{RB}_h^k(s,a)$ and the bonus $b_h^k(s,a)$ as

$$\mathrm{RB}_h^k(s,a) = \widehat{r}_h^k(s,a) - \beta \ln \mathbb{E}_{\widehat{P}_h^k}\big[\exp\big(-\beta^{-1}V_{h+1}^{k,\beta,\eta}\big)\big](s,a),$$
$$b_h^k(s,a) = \big(1+\beta e^{\beta^{-1}H}\sqrt{S}\big)\sqrt{\frac{2\ln(2SAHK/\delta)}{n_h^{k-1}(s,a)\vee 1}}.$$

Tabular-χ^2 Setting According to the dual formulation for regularized χ^2 setting in Lemma C.9, given the estimated value function $V_{h+1}^{k,\beta,\eta}$ and empirical reward $\widehat{r}_h^k$ and transition $\widehat{P}_h^k$, we choose the robust Bellman estimator $\mathrm{RB}_h^k(s,a)$ and the bonus $b_h^k(s,a)$ as

$$\mathrm{RB}_h^k(s,a) = \widehat{r}_h^k(s,a) + \sup_{\boldsymbol{\lambda}\in[0,H]}\big(\mathbb{E}_{\widehat{P}_h^k}\big[V_{h+1}^{k,\beta,\eta}-\boldsymbol{\lambda}\big](s,a) - \frac{1}{4\beta}\mathrm{Var}_{\widehat{P}_h^k}\big[V_{h+1}^{k,\beta,\eta}-\boldsymbol{\lambda}\big](s,a)\big),$$
$$b_h^k(s,a) = \left(2+\frac{3H}{4\beta}\right)H\sqrt{\frac{2S^2\ln(48SAH^3K^2/\delta)}{n_h^{k-1}(s,a)\vee 1}} + \frac{1+4\beta}{4\beta K}.$$

4.3 Policy Update

For the selection of policy π_h^k, $(h,k)\in[H]\times[K]$, we first present a generic form (4.3), which maximizes the estimated value function $V_h^k(s)$.

$$\begin{aligned}\pi_h^k(\cdot|s) &= \operatorname*{argmax}_{\pi\in\Pi}\big\{\mathbb{E}_{\pi(\cdot|s)}\big[Q_h^{k,\beta,\eta}(s,\cdot)\big] - \Omega_s^\eta(\pi)\big\}\\ &= \operatorname*{argmax}_{\pi\in\Pi}\big\{\mathbb{E}_{\pi(\cdot|s)}\big[Q_h^{k,\beta,\eta}(s,\cdot)\big] - \sigma_{\mathcal{R}_s^\eta(\pi)}(-\pi(\cdot|s))\big\},\end{aligned} \tag{4.3}$$

where we specify the choice of policy regularization according to Proposition 3.2 in the second equality.

Next, for different types of the policy regularization, we specify the corresponding policy update formulation. The exact form of the reward-uncertainty interval for the associated support function is established in Derman et al. (2021, Section 3).

Negative Shannon Entropy Let $\mathcal{R}_{s,a}^{\mathrm{NS},\eta}(\pi) = \big[\eta\ln(1/\pi(a|s)),+\infty\big)$ for all $(\pi,s,a)\in\Pi\times\mathcal{S}\times\mathcal{A}$, and the associated support function is

$$\sigma_{\mathcal{R}_s^{\mathrm{NS},\eta}(\pi)}\big(-\pi(\cdot|s)\big) = \eta\cdot\sum_{a\in\mathcal{A}}\pi(a|s)\ln\big(\pi(a|s)\big).$$

For this setting, we update the policy as

$$\pi_h^k(\cdot|s)\propto\exp\big(Q_h^{k,\beta,\eta}(s,\cdot)/\eta\big). \tag{4.4}$$

Kullback-Leibler Divergence Fixing a reference policy $\widetilde{\pi}$, let $\mathcal{R}_{s,a}^{\mathrm{KL},\eta}(\pi) = \eta \ln\big(\widetilde{\pi}(a|s)\big) + \mathcal{R}_{s,a}^{\mathrm{NS},\eta}(\pi)$ for all $(\pi, s, a) \in \Pi \times \mathcal{S} \times \mathcal{A}$, and the associated support function is

$$\sigma_{\mathcal{R}_s^{\mathrm{KL},\eta}(\pi)}\big(-\pi(\cdot|s)\big) = \eta \cdot \sum_{a\in\mathcal{A}} \pi(a|s) \ln\big(\pi(a|s)/\widetilde{\pi}(a|s)\big).$$

For this setting, we update the policy as

$$\pi_h^k(\cdot|s) \propto \widetilde{\pi}_h^k(\cdot|s) \exp\big(Q_h^{k,\beta,\eta}(s,\cdot)/\eta\big). \tag{4.5}$$

Negative Tsallis Entropy Let $\mathcal{R}_{s,a}^{\mathrm{NT},\eta}(\pi) = \eta[(1-\pi(a|s))/2, +\infty)$ for all $(\pi, s, a) \in \Pi \times \mathcal{S} \times \mathcal{A}$, and the associated support function is

$$\sigma_{\mathcal{R}_s^{\mathrm{NT},\eta}(\pi)}\big(-\pi(\cdot|s)\big) = \eta \cdot (\|\pi(\cdot|s)\|^2 - 1)/2.$$

For this setting, we update the policy as

$$\pi_h^k(\cdot|s) = \mathrm{proj}_\Delta\big(Q_h^{k,\beta,\eta}(s,\cdot)/\eta\big). \tag{4.6}$$

In Lu et al. (2024); Liu and Xu (2024a); He et al. (2025), the updated policy is chosen as the greedy policy with respect to the estimated Q-function $Q_h^{k,\beta,\eta}$. In contrast, since our formulation incorporates a policy regularization term, the corresponding robust Bellman equations Proposition 6.2 differ from theirs, leading to a different policy update rule. Notably, the formulation in (4.3) allows us to employ a soft policy update.

5 THEORETICAL RESULTS

In this section, we present the regret bounds for our proposed algorithm RSPVI. We begin by introducing the assumptions required for online robust learning. Then we discuss our results under different f-divergence settings.

He et al. (2025) showed that online robust learning can be exponentially hard without appropriate assumptions. Based on this, they introduce the bounded visitation measure ratio assumption and justify its necessity by deriving the lower bounds. Following their approach, we begin by introducing several notations that will be useful throughout the discussion.

Definition 5.1 (Worst-case transition)**.** In the tabular setting, for any policy π, we define the worst-case transition corresponding to π as

$$P_h^{w,\pi}(\cdot|s,a) = \operatorname*{argmin}_{P_h\in\Delta(S)} \mathbb{E}_{P_h}[V_{h+1}^{\pi,\beta,\eta}](s,a) + \beta \cdot \mathrm{D}(P_h(\cdot|s,a)\|P_h^o(\cdot|s,a)).$$

Based on Definition 5.1, we also define the corresponding visitation measure.

Definition 5.2 (Visitation measure)**.** At timestep $h \in [H]$, we denote $\mathrm{d}_h^\pi(\cdot)$ as the visitation measure on $\mathcal{S}$ induced by policy π under P^o, and $\mathrm{q}_h^\pi(\cdot)$ as the visitation measure on $\mathcal{S}$ induced by policy π under $P^{w,\pi}$.

We now make the same bounded visitation measure ratio assumption as in He et al. (2025).

Assumption 5.3 (Bounded visitation measure ratio)**.** Under the definition of Definition 5.2, we define $C_{vr} := \sup_{\pi,h,s} \frac{\mathrm{q}_h^\pi(s)}{\mathrm{d}_h^\pi(s)}$ as the supremal ratio between the nominal visitation measure and the worst-case visitation measure. We assume that C_{vr} is polynomial in H, S and A.

Theorem 5.4 (Tabular RRMDP regret bounds)**.** Assume that Assumption 5.3 holds for each f-divergence-based transition regularization, and consider the policy regularizations in (4.4), (4.5), and (4.6) Then for any $\delta \in (0, 1/3)$, with probability at least $1 - 3\delta$, the regret of Algorithm 1 satisfies

$$\mathrm{Regret}(K) = \begin{cases} \widetilde{\mathcal{O}}\big(C_{vr} S^{\frac{3}{2}} A H^2 + C_{vr}^{\frac{1}{2}} S A^{\frac{1}{2}} H^2 \sqrt{K}\big) & \text{(TV)} \\ \widetilde{\mathcal{O}}\big(\big(1 + \beta e^{\beta^{-1} H} \sqrt{S}\big)\big(C_{vr} S A H + C_{vr}^{\frac{1}{2}} S^{\frac{1}{2}} A^{\frac{1}{2}} H \sqrt{K}\big)\big) & \text{(KL)} \\ \widetilde{\mathcal{O}}\big(C_{vr} S^2 A H^3 + C_{vr}^{\frac{1}{2}} S^{\frac{3}{2}} A^{\frac{1}{2}} H^3 \sqrt{K}\big) & (\chi^2) \end{cases}.$$

Theorem 5.4 presents the first regret bound for online MDPs that simultaneously accounts for both transition and reward uncertainties, demonstrating that sample-efficient online robust learning is achievable under Assumption 5.3. Unlike transition dynamics, we make no assumptions on reward uncertainty, since reward variations do not affect the difficulty of reaching specific states across the source and target environments. A key distinction of our setting is that the optimal policy is soft, whereas in He et al. (2025) it must be greedy. This advantage arises from policy regularization, and often leads to the learned policy more exploratory and reliable.

The orders in Theorem 5.4 match those in He et al. (2025), showing that incorporating an additional mechanism for reward robustness does not increase the overall regret. This highlights the sample efficiency of RSPVI and the advantages of our proposed doubly regularized framework. The key insight is as follows: in the online setting, we add a bonus to the estimated function to ensure optimism, and the estimation error can be bounded by this bonus term. Thanks to our algorithmic design, the optimism property continues to hold for all estimated value functions and Q-functions when the bonus terms are chosen in the same way as before. Consequently, by applying Assumption 5.3 in the same manner as He et al. (2025), the order of the regret remains unchanged.

6 LINEAR MDP EXTENSION

6.1 Preliminaries

Linear Function Approximation d-rectangular linear regularized robust MDP was first proposed in Tang et al. (2025), which changes the uncertainty set into a penalty term while persisting the linear structure of the transitions and reward functions. We make the following assumption the same as Liu and Xu (2024a); Tang et al. (2025).

Assumption 6.1. (Jin et al., 2020) Given a known state-action feature mapping $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ satisfying $\sum_{i=1}^d \phi_i(s,a) = 1, \phi_i(s,a) \geq 0$. Further more, we assume the reward function $\{r_h\}_{h=1}^H$ and the nominal transition kernels $\{P_h^o\}_{h=1}^H$ admit linear structures. Specifically, we have for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$,

$$r_h(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h \rangle, \quad P_h^o(\cdot|s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h^o(\cdot) \rangle,$$

where $\{\boldsymbol{\theta}_h\}_{h=1}^H$ are unknown vectors with bounded norm $\|\boldsymbol{\theta}_h\|_2 \leq \sqrt{d}$ and $\{\boldsymbol{\mu}_h^o\}_{h=1}^H$ are unknown probability measure vectors over $\mathcal{S}$, i.e., $\boldsymbol{\mu}_h^o = (\mu_{h,1}^o, \mu_{h,2}^o, \cdots, \mu_{h,d}^o)$, $\mu_{h,i}^o \in \Delta(\mathcal{S}), \forall\, i \in [d]$.

Based on such linear structure, the robust value function and Q-function are defined as

$$V_h^{\pi,\beta,\eta}(s) = \inf_{\substack{\boldsymbol{\mu}_t \in \Delta(S)^d \\ P_t = \langle \boldsymbol{\phi}, \boldsymbol{\mu}_t \rangle}} \mathbb{E}_{\pi,\{P_t\}_{t=h}^H} \Bigg[\sum_{t=h}^H \Big(r_t(s_t, a_t) + \beta \cdot \big\langle \boldsymbol{\phi}(s_t, a_t), \mathbf{D}\big(\boldsymbol{\mu}_t \| \boldsymbol{\mu}_t^o\big) \big\rangle - \Omega_{s_t}^\eta(\pi_t) \Big) \,\Big|\, s_h = s \Bigg],$$

$$Q_h^{\pi,\beta,\eta}(s, a) = \inf_{\substack{\boldsymbol{\mu}_t \in \Delta(S)^d \\ P_t = \langle \boldsymbol{\phi}, \boldsymbol{\mu}_t \rangle}} \mathbb{E}_{\pi,\{P_t\}_{t=h}^H} \Bigg[\sum_{t=h}^H \Big(r_t(s_t, a_t) + \beta \cdot \big\langle \boldsymbol{\phi}(s_t, a_t), \mathbf{D}\big(\boldsymbol{\mu}_t \| \boldsymbol{\mu}_t^o\big) \big\rangle \Big) - \sum_{t=h+1}^H \Omega_{s_t}^\eta(\pi_t) \,\Big|\, s_h = s, a_h = a \Bigg],$$

where $\mathbf{D}(\boldsymbol{\mu} \| \boldsymbol{\mu}^o) = (\mathrm{D}(\mu_1 \| \mu_1^o), \cdots, \mathrm{D}(\mu_d \| \mu_d^o))^\top$.

It is worth noting that there has been no prior framework addressing reward robustness in the linear setting, largely due to the constraints of its linear structure. In contrast, under our doubly regularized framework, reward robustness naturally emerges from the policy regularization, making such robustness much easier to achieve.

Dynamic Programming Principle Similar to the tabular setting, we have the dynamic programming principle for doubly regularized MDPs as follows

Proposition 6.2. For doubly regularized linear MDPs, it holds that for any policy π and any $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$,

$$Q_h^{\pi,\beta,\eta}(s, a) = \inf_{\substack{\boldsymbol{\mu}_h \in \Delta(S)^d \\ P_h = \langle \boldsymbol{\phi}, \boldsymbol{\mu}_h \rangle}} \big\{ \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[V_{h+1}^{\pi,\beta,\eta}(s')] + \beta \cdot \big\langle \boldsymbol{\phi}(s, a), \mathbf{D}\big(\boldsymbol{\mu}_h \| \boldsymbol{\mu}_h^o\big) \big\rangle \big\} + r_h(s, a), \tag{6.1}$$

$$V_h^{\pi,\beta,\eta}(s) = -\Omega_s^\eta(\pi_h) + \mathbb{E}_{a \sim \pi_h(\cdot|s)}\big[Q_h^{\pi,\beta,\eta}(s, a)\big]. \tag{6.2}$$

6.2 Algorithm

Then, we present the linear version of Algorithm 1 as Algorithm 2.

Q-function Estimation For the estimation of robust Q-function $Q_h^{k,\beta,\eta}$, $(h, k) \in [H] \times [K]$, we first present a generic form (6.3). Since the feature mapping is known, it is sufficient to estimate the weight vectors $\boldsymbol{\theta}_h^k$ and $\boldsymbol{w}_h^k$ to recover the robust Q-function. Here, $\boldsymbol{\theta}_h^k$ corresponds to the reward component, and $\boldsymbol{w}_h^k$ corresponds to the estimated value function $V_{h+1}^{k,\beta,\eta}$. Instead of a model-based method in the tabular setting, we use ridge regression to estimate $\boldsymbol{\theta}_h^k$ and $\boldsymbol{w}_h^k$. $\Gamma_h^k(s, a)$ denotes the reward term, where c needs to be determined according to specific f-divergence setting. In the following, we instantiate this generic form under different f-divergence settings and provide explicit formulations for robust Bellman estimation as well as the coefficient c for the bonus term.

$$Q_h^k(s, a) = \min\big\{ \big\langle \boldsymbol{\phi}(s, a), \boldsymbol{\theta}_h^k + \boldsymbol{w}_h^k \big\rangle + \Gamma_h^k(s, a), H - h + 1 \big\}, \tag{6.3}$$

$$\Gamma_h^k(s, a) = c \sum_{i=1}^d \|\phi_i(s, a) \mathbb{1}_i\|_{(\Lambda_h^k)^{-1}},$$

Algorithm 2 RSPVI for linear setting

Require: transition regularizer $\beta > 0$, policy regularizer Ω^η, ridge regression regularizer $\lambda > 0$.

for $k = 1, \cdots, K$ **do**
  $V_{H+1}^{k,\beta,\eta}(\cdot) \leftarrow 0$.
  for $h = H, \cdots, 1$ **do**
    for $\forall\, (s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
      Update Q-function estimation $Q_h^{k,\beta,\eta}(s,a)$ according to (6.3).
    end for
    for $\forall\, s \in \mathcal{S}$ **do**
      Update policy $\pi_h^k(\cdot|s)$ according to (4.3).
      Update value function estimation $V_h^{k,\beta,\eta}(s) = \big\langle Q_h^{k,\beta,\eta}(s,\cdot), \pi_h^k(\cdot|s)\big\rangle - \Omega_s^\eta(\pi_h^k)$.
    end for
  end for
  Collect a trajectory τ^k by executing π^k.
end for

where $\Lambda_h^k = \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau)\boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top + \lambda I$, $\lambda > 0$ is the regularizer in the ridge regression.

Linear-TV Setting According to the linear structure assumption in Assumption 6.1 and the dual formulation for regularized TV setting in Lemma D.3, given the estimated robust value function $V_{h+1}^{k,\beta,\eta}$, we estimate the parameters $\boldsymbol{\theta}_h^k$ and $\boldsymbol{w}_h^k$ as follows

$$\boldsymbol{\theta}_h^k = \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^d} \sum_{\tau=1}^{k} \big(r_h^\tau - \boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top \boldsymbol{\theta}\big)^2 + \lambda\|\boldsymbol{\theta}\|_2^2,$$

$$\boldsymbol{w}_h^k = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \sum_{\tau=1}^{k} \Big(\big[V_{h+1}^{k,\beta,\eta}(s_{h+1}^\tau)\big]_{\min_{s'\in\mathcal{S}} V_{h+1}^{k,\beta,\eta}(s')+\beta} - \boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top \boldsymbol{w}\Big)^2 + \lambda\|\boldsymbol{w}\|_2^2.$$

The bonus coefficient is given by $c = Hd \cdot \xi_{\mathrm{TV}}$, where

$$\xi_{\mathrm{TV}} = 720 + 3\sqrt{40 \log\big(96K^{13/2}H|\mathcal{A}|^3/\delta\big)}.$$

Linear-KL Setting According to the linear structure assumption in Assumption 6.1 and the dual formulation for regularized KL setting in Lemma D.6, we estimate $\boldsymbol{w}_h^k$ through two steps. First, given the estimated robust value function $V_{h+1}^{k,\beta,\eta}$, we estimate $\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}]$ as

$$\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}] = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \sum_{\tau=1}^{k} \big(e^{-V_{h+1}^{k,\beta,\eta}(s_{h+1}^\tau)/\beta} - \boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top \boldsymbol{w}\big)^2 + \lambda\|\boldsymbol{w}\|_2^2.$$

And then we estimate the parameters $\boldsymbol{\theta}_h^k$ and $\boldsymbol{w}_h^k$ as

$$\boldsymbol{\theta}_h^k = \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^d} \sum_{\tau=1}^{k} \big(r_h^\tau - \boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top \boldsymbol{\theta}\big)^2 + \lambda\|\boldsymbol{\theta}\|_2^2,$$

$$\boldsymbol{w}_h^k = -\beta \log \max\big\{\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}], e^{-H/\beta}\big\}.$$

The truncation here is to make V_h^k lower bounded and therefore the ridge regression operation is well-defined. And the bonus coefficient is given by $c = (1 + 2\beta e^{\beta^{-1}H})Hd \cdot \xi_{\mathrm{KL}}$, where

$$\xi_{\mathrm{KL}} = 80 + \sqrt{40 \log\big(64\beta K^{11/2} H d^{1/2}(1 + 2\beta e^{\beta^{-1}H})^2 |\mathcal{A}|^3/\delta\big)}.$$

Linear-χ^2 Setting According to the linear structure assumption in Assumption 6.1 and the dual formulation for regularized χ^2 setting in Lemma D.9, we estimate $\boldsymbol{w}_h^k$ through two steps. First, given the estimated robust value function $V_{h+1}^{k,\beta,\eta}$ and an arbitrary constant α, we estimate $\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}(s)]_\alpha$ and $\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}(s)]_\alpha^2$ as

$$\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}(s)]_\alpha = \left[\operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \sum_{\tau=1}^{k} \big([V_{h+1}^{k,\beta,\eta}(s_{h+1}^\tau)]_\alpha - \boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top \boldsymbol{w}\big)^2 + \lambda\|\boldsymbol{w}\|_2^2\right]_{[0,H]},$$

$$\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}(s)]_\alpha^2 = \left[\operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \sum_{\tau=1}^{k} \big([V_{h+1}^{k,\beta,\eta}(s_{h+1}^\tau)]_\alpha^2 - \boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top \boldsymbol{w}\big)^2 + \lambda\|\boldsymbol{w}\|_2^2\right]_{[0,H]}.$$

Based on the estimated expectations and the dual variable α, we estimate the parameters $\boldsymbol{\theta}_h^k$ and $\boldsymbol{w}_h^k$ as

$$\begin{aligned}
\boldsymbol{\theta}_h^k &= \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^d} \sum_{\tau=1}^{k} \big(r_h^\tau - \boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top \boldsymbol{\theta}\big)^2 + \lambda\|\boldsymbol{\theta}\|_2^2, \\
\boldsymbol{w}_h^k &= \sup_{\alpha\in[0,H]} \left\{\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}(s)]_\alpha - \frac{1}{4\beta}\widehat{\mathrm{Var}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}(s)]_\alpha\right\} \\
&= \sup_{\alpha\in[0,H]} \left\{\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}(s)]_\alpha + \frac{1}{4\beta}\big(\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}(s)]_\alpha\big)^2 - \frac{1}{4\beta}\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}(s)]_\alpha^2\right\}.
\end{aligned}$$

The bonus coefficient is given by $c = \big(1 + H/(2\beta)\big)Hd \cdot \xi_{\chi^2}$, where

$$\xi_{\chi^2} = 720 + 3\sqrt{40 \log\big(96K^6H^5(1 + H/(2\beta))^3|\mathcal{A}|^3/\delta\big)}.$$

6.3 Theoretical Results

As in the tabular setting, we first introduce the necessary definitions and assumptions.

Definition 6.3 (Worst-case transition)**.** In the linear setting, since the transition model $P_h^o(\cdot|s,a) = \langle\boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h^o(\cdot)\rangle$ is a factor distribution here, for any policy π, we first define $\boldsymbol{\mu}_h^{w,\pi} = \big(\mu_{h,1}^{w,\pi}, \cdots, \mu_{h,d}^{w,\pi}\big)^\top$, where for any $i \in [d]$,

$$\mu_{h,i}^{w,\pi}(\cdot) = \operatorname*{argmin}_{\mu_h\in\Delta(S)} \mathbb{E}_{\mu_h}[V_{h+1}^{\pi,\beta,\eta}](s,a) + \beta \cdot \mathrm{D}(\mu_h(\cdot)\|\mu_{h,i}^o(\cdot)).$$

Then we can directly obtain the worst-case transition from the linear structure assumption Assumption 6.1, written as

$$P_h^{w,\pi}(\cdot|s,a) = \big\langle\boldsymbol{\phi}(s,a), \mu_{h,i}^{w,\pi}(\cdot)\big\rangle.$$

We define the worst-case transition differently in the tabular and linear settings, since in the linear case the transitions additionally admit a linear structure. With such definition in place, we can now naturally define the corresponding visitation measures as follows

Definition 6.4 (Visitation measure)**.** At timestep $h \in [H]$, we denote $\mathrm{d}_h^\pi(\cdot)$ as the visitation measure on $\mathcal{S}$ induced by policy π under P^o, and $\mathrm{q}_h^\pi(\cdot)$ as the visitation measure on $\mathcal{S}$ induced by policy π under $P^{w,\pi}$.

We now introduce the linear version of the bounded visitation measure ratio assumption as follows

Assumption 6.5 (Bounded visitation measure ratio)**.** Under the definition of Definition 6.4, we define $C_{vr} := \sup_{\pi,h,s} \frac{\mathrm{q}_h^\pi(s)}{\mathrm{d}_h^\pi(s)}$ as the supremal ratio between the nominal visitation measure and the worst-case visitation measure. We assume that C_{vr} is polynomial in H, S and A.

Note that the formulations in Definition 6.4 and Assumption 6.5 are the same as those in Definition 5.2 and Assumption 5.3 in the tabular setting. This showcases that C_{vr} is an intrinsic quantity characterizing the difficulty of the online robust learning problem, independent of the linear structure assumption. Next, we present our theoretical results in the linear setting. As in Liu and Xu (2024a), we begin by presenting the instance-dependent upper bounds.

Theorem 6.6. Assume Assumption 6.1 and Assumption 6.5 holds for each f-divergence-based transition regularization, and consider the policy regularizations in (4.4), (4.5), and (4.6). Then by choosing $\lambda = 1$, for any $\delta \in (0, 1/3)$, with probability at least $1 - 3\delta$, the regret of Algorithm 2 satisfies

$$\mathrm{Regret}(K) \le 2C_{vr}\left(\sqrt{2KH^3\ln(2/\delta)} + c\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{d}\left\|\phi_i(s_h^k, a_h^k)\mathbb{1}_i\right\|_{(\Lambda_h^k)^{-1}}\right), \tag{6.4}$$

where c is the parameter in bonus coefficients defined in Section 6.2 for each f-divergence setting.

Theorem 6.6 provides the first finite-sample regret bound in both online distributionally robust learning under general f-divergences with linear function approximation and reward-robust learning. This result demonstrates that when Assumption 6.5 is satisfied, efficient online robust learning is attainable. The formulation of (6.4) is similar to that of Liu and Xu (2024a). However, since we impose the bounded visitation measure assumption rather than the fail-state assumption (Liu and Xu, 2024a, Assumption 4.1), the bound includes an additional dependence on C_{vr}.

Theorem 6.7. Assume Assumption 6.1 holds for each f-divergence-based transition regularization, and consider the policy regularizations in (4.4), (4.5), and (4.6). Instead of assuming Assumption 6.5, we now assume that

$$\mathbb{E}_\pi\left[\boldsymbol{\phi}(s_h, a_h)\boldsymbol{\phi}(s_h, a_h)^\top\right] \ge \alpha I \tag{6.5}$$

for all $(\pi, h) \in \Pi \times [H]$, where $\alpha > 0$ is a constant. Then by choosing $\lambda = 1$, for any $\delta \in (0, 1/3)$, with probability at least $1 - 3\delta$, the regret of Algorithm 2 satisfies

$$\mathrm{Regret}(K) \le 2cH\sqrt{K}\sqrt{\frac{128}{\alpha^2}\log\left(\frac{dHK}{\delta}\right) + \frac{2}{\alpha}\log(K)},$$

where c is the parameter in bonus coefficients defined in Section 6.2 for each f-divergence setting.

As discussed in Liu and Xu (2024a), the d-rectangular estimation error in (6.4), which arises from the structure of the d-rectangular uncertainty set, cannot be bounded directly using the elliptical potential lemma (Abbasi-Yadkori et al., 2011). To address this difficulty, Liu and Xu (2024a, Corollary 5.3) introduce an additional (6.5) assumption. When specialized to the tabular setting, $\boldsymbol{\phi}(s_h, a_h)$ reduces to a one-hot vector. In this case, the diagonal entry of the matrix $\mathbb{E}_\pi[\boldsymbol{\phi}(s_h, a_h)\boldsymbol{\phi}(s_h, a_h)^\top]$ corresponds to the probability of visiting $(s, a) \in \mathcal{S} \times \mathcal{A}$ at step h under policy π. Hence, we have $\mathrm{d}_h^\pi(s, a) \geq \alpha$ for all $(\pi, h, s, a) \in \Pi \times [H] \times \mathcal{S} \times \mathcal{A}$, which directly implies Assumption 5.3. In the linear setting, (6.5) can be interpreted as a uniform coverage condition, requiring the nominal environment to be sufficiently exploratory. Theorem 6.7 shows that (6.5) can also serve as a sufficient condition for online robust learning.

Compared with Liu and Xu (2024a, Corollary 5.3), their result additionally relies on a fail-state assumption (Liu and Xu, 2024a, Assumption 4.1) in addition to (6.5). Moreover, Liu and Xu (2024a, Corollary 5.3) is restricted to transition uncertainty characterized by Total Variance, whereas Theorem 6.7 extends to more general f-divergence-based transition robustness.

7 EXPERIMENT

In this section, we validate our theoretical findings through an empirical study. It is worth noting that this is the first attempt in literature to make linear MDP practical in the context of robust RL. We choose the standard CartPole environment as a start, where the objective is to keep a pole balanced by moving a cart left or right. An episode is terminated if the pole falls or is truncated after 200 steps. The state space is continuous and represented by the tuple $(x, \dot{x}, \theta, \dot{\theta})$, corresponding to the cart's position, velocity, the pole's angle, and angular velocity, respectively. The action space is discrete with two possible actions, $0, 1$, indicating the direction of force applied by the agent at each step. As long as the episode continues, the agent receives a reward of $r = 1$ at each time step.

To implement RSPVI in the linear setting, we initially learn a feature mapping ϕ using a discrete variational autoencoder (VAE) (Rolfe, 2016). The dimension of the feature mapping ϕ is a hyperparameter that we tune from 5 to 100 in our experiment and select the one ($d = 80$ in this CartPole environment) that effectively models the dynamics. Given the high feature dimension, we observe that computing the bonus in Algorithm 2 incurs significant computational overhead and becomes the primary bottleneck of the algorithm's speed. Consequently, we opt not to include it in the implementation, albeit at the potential cost of slight performance degradation. During evaluation, we introduce perturbation to the agent by randomly selecting an action with a specified probability at each step. The learned policy is subsequently assessed by averaging the cumulative reward across 50 runs, and all results are further averaged across 10 random seeds.

Experiment results are shown in Figure 1, for which we choose TV and KL as the transition regularization and Negative Shannon Entropy as the policy regularization. As we can see from Figure 1, our algorithm RSPVI is less sensitive to the perturbation compared with the non-robust baseline. We also compare RSPVI with DR-LSVI-UCB (Liu and Xu, 2024a). We find that DR-LSVI-UCB fails to learn the robust policy without the exploration bonus, while the soft policies learned by RSPVI has exploration ability in nature. This further demonstrates the advantages of a soft policy and our doubly regularized framework in applications with large state-action spaces. More experimental results are provided in the appendix.

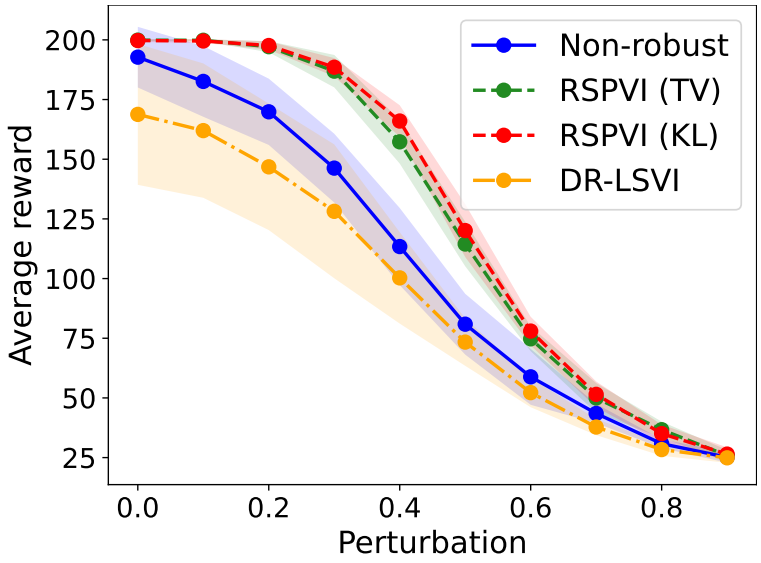


Figure 1: Experiment results for the CartPole environment. RSPVI is equipped with Negative Shannon entropy regularization. x-axis represents the perturbation level during evaluation. y-axis represents the average cumulative reward achieved over 50 episodes.

8 CONCLUSION

We propose a novel doubly regularized framework that achieves both reward robustness and transition robustness. Within this framework, we introduce an online learning algorithm, RSPVI, and prove that when using general f-divergence-based transition regularization together with various policy regularization formulations, the sample complexity remains no worse than that of traditional RMDPs. We further extend RSPVI to the linear setting, establishing the first finite-sample regret guarantee under general f-divergence. Compared with prior work, our results require fewer structural assumptions. Finally, we demonstrate the effectiveness and robustness of our framework through experiments in the CartPole environment.

A Additional Details on Experiments

In this section, we provide more experimental results and details in addition to Section 7. First, we report the hyper-parameters used in the experiments in Table 1. In practice, we find that the ridge regression coefficient λ plays a crucial role in training and is tuned from $\{1, 0.1, 0.01, 0.001, 0.0001\}$.

First, we conduct ablation study to show the impact of η on the algorithm's performance with negative Shannon entropy policy regularization.

Negative Shannon Entropy As shown in Figures 2(a) and 2(d), it is clear that selecting an appropriate value for η is crucial for algorithm training. When η is too large, the excessive penalty on policy regularization can drive the policy toward near-random behavior, preventing the agent from gathering sufficient useful information during exploration. Conversely, when η is too small, the policy becomes less stochastic, which reduces exploration and may lead to the agent learning a suboptimal policy.

Since our theoretical framework also covers Kullback-Leibler divergence and negative Tsallis entropy policy regularization, we conduct additional experiments to evaluate the performance of RSPVI under these two settings.

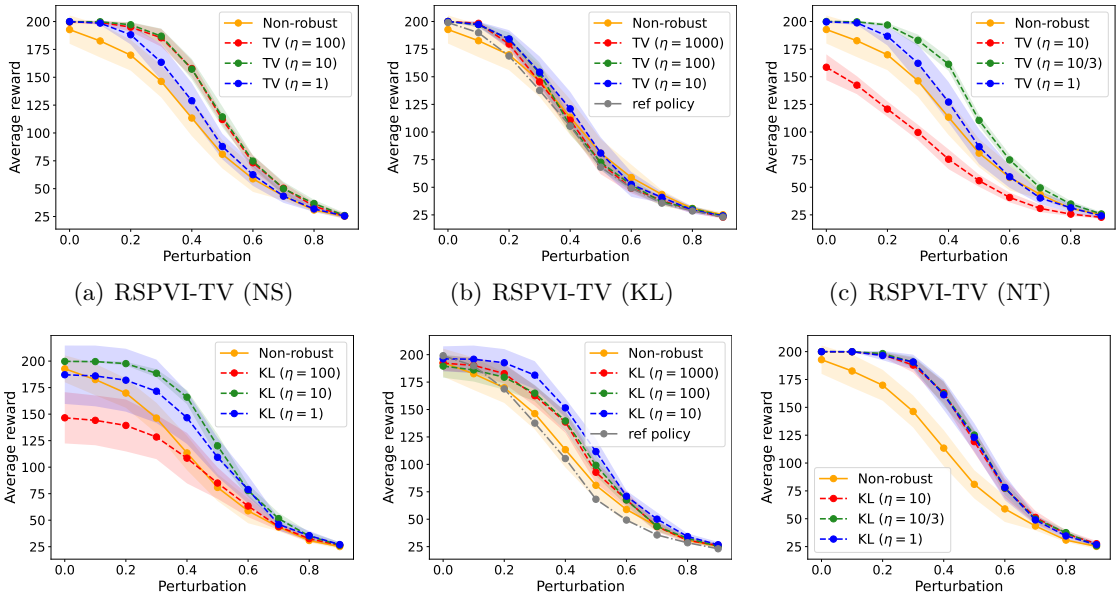
Kullback-Leibler Divergence In Figures 2(b) and 2(e), we present the results of RSPVI with KL divergence policy regularization for different values of η. The reference policy is learned with NS policy regularization but without transition regularization, making it non-robust to environmental perturbations. As η increases, the stronger policy regularization brings the learned policy closer to the reference policy. However, the overall performance of the learned policy still surpasses that of the reference policy. Further, the KL-divergence defined transition regularization leads to better robustness performance compared to the TV case.

Negative Tsallis Entropy In Figures 2(c) and 2(f), we present the results of RSPVI with negative Tsallis entropy policy regularization for different values of η. With TV-distance defined transition regularization, similar to the behavior observed with Negative Shannon Entropy, the policy can be too random for large η and becomes too deterministic for small η, while a moderate η lead to the optimal performance. While for the case with KL-divergence defined transition regularization, our algorithm is not sensitive to the choice of η.

In general, this CartPole experiment suggests that our robust algorithm RSPVI outperforms the non-robust baseline.

Table 1: hyper-parameters for experiments

Symbol	Description	Value
d	latent space dimension	80
K	total episodes	2000
λ	ridge regression parameter	0.001
γ	discount factor	0.99
β_{TV}	transition regularization parameter	70
β_{KL}	transition regularization parameter	120



(a) RSPVI-TV (NS) (b) RSPVI-TV (KL) (c) RSPVI-TV (NT)

(d) RSPVI-KL (NS) (e) RSPVI-KL (KL) (f) RSPVI-KL (NT)

Figure 2: Performance of RSPVI with different types of policy regularization and values of η. Figures 2(a) to 2(c) use TV transition regularization, while Figures 2(d) to 2(f) use KL transition regularization. Additionally, Figures 2(a) and 2(d) apply negative Shannon entropy (NS) policy regularization, Figures 2(b) and 2(e) apply Kullback-Leibler divergence (KL) policy regularization, and Figures 2(c) and 2(f) apply negative Tsallis entropy (NT) policy regularization. The x-axis represents the perturbation level during evaluation, and the y-axis represents the average cumulative reward achieved over 50 episodes.

B Proof of the Policy Update Formulation

In particular, we prove that the policy update formulation in (4.4), (4.5), (4.6) satisfies (4.3).

Proof of the Kullback-Leibler Divergence case. Fixing a reference policy $\widetilde{\pi}$, we need to prove that policy

$$\pi(\cdot|s) \propto \widetilde{\pi}_h^k(\cdot|s) \exp\left(Q_h^{k,\beta,\eta}(s,\cdot)/\eta\right)$$

maximizes the objective $J(\pi)$

$$J(\pi) = \left\langle Q_h^{k,\beta,\eta}(s,\cdot), \pi(\cdot|s)\right\rangle - \eta \cdot \sum_{a\in\mathcal{A}} \pi(a|s) \ln\left(\frac{\pi(a|s)}{\widetilde{\pi}_h^k(a|s)}\right).$$

In order to prove this, we define an auxiliary policy π' as

$$\pi'(a|s) = \frac{\widetilde{\pi}_h^k(a|s) \exp\left(Q_h^{k,\beta,\eta}(s,a)/\eta\right)}{Z}, \quad \text{where} \quad Z = \sum_a \widetilde{\pi}_h^k(a|s) \exp\left(Q_h^{k,\beta,\eta}(s,a)/\eta\right).$$

From the non-negativity of KL divergence, we have

$$\begin{aligned}0 &\leq \sum_a \pi(a|s) \log\left(\frac{\pi(a|s)}{\pi'(a|s)}\right) \\ &= \sum_a \pi(a|s) \log\left(\frac{\pi(a|s)}{\widetilde{\pi}_h^k(a|s)}\right) - \frac{1}{\eta}\sum_a \pi(a|s) Q_h^{k,\beta,\eta}(s,a) + \log Z.\end{aligned}$$

Rearrange this and it follows that

$$\sum_a \pi(a|s) Q_h^{k,\beta,\eta}(s,a) - \eta \sum_a \pi(a|s) \log\left(\frac{\pi(a|s)}{\widetilde{\pi}_h^k(a|s)}\right) \leq \eta \log Z.$$

The equality is achieved if and only if $\mathrm{KL}(\pi\|\pi') = 0$, or $\pi = \pi'$. Since $J(\pi)$ is strictly concave with respect to π, π' is the unique maximizer of $J(\pi)$. □

Proof of the Negative Shannon Entropy case. We need to prove that policy

$$\pi(\cdot|s) \propto \exp\big(Q_h^{k,\beta,\eta}(s,\cdot)/\eta\big)$$

maximizes the objective $J(\pi)$

$$J(\pi) = \big\langle Q_h^{k,\beta,\eta}(s,\cdot), \pi(\cdot|s)\big\rangle - \eta \cdot \sum_{a\in\mathcal{A}} \pi(a|s) \ln\big(\pi(a|s)\big).$$

The result is directly followed by the Kullback-Leibler Divergence case by choosing $\widetilde{\pi}(\cdot|s)$ as the uniform distribution. □

Proof of the Negative Tsallis Entropy case. We need to prove that policy

$$\pi(\cdot|s) = \mathrm{proj}_\Delta\big(Q_h^{k,\beta,\eta}(s,\cdot)/\eta\big)$$

maximizes the objective $J(\pi)$

$$J(\pi) = \big\langle Q_h^{k,\beta,\eta}(s,\cdot), \pi(\cdot|s)\big\rangle - \frac{\eta}{2}\|\pi(\cdot|s)\|^2,$$

In order to prove this, we rewrite the objective as

$$\begin{aligned}J(\pi) &= \big\langle Q_h^{k,\beta,\eta}(s,\cdot), \pi(\cdot|s)\big\rangle - \frac{\eta}{2}\|\pi(\cdot|s)\|^2 \\ &= -\frac{\eta}{2}\left\|\pi(\cdot|s) - \frac{1}{\eta}Q_h^{k,\beta,\eta}(s,\cdot)\right\|^2 + \frac{1}{2\eta}\|Q_h^{k,\beta,\eta}(s,\cdot)\|^2.\end{aligned}$$

Since $\|Q_h^{k,\beta,\eta}(s,\cdot)\|^2/(2\eta)$ is independent of π, it is equivalent to minimize $\|\pi(\cdot|s) - Q_h^{k,\beta,\eta}(s,\cdot)/\eta\|^2$, subject to $\pi(\cdot|s)$ being a distribution. As the projection operator proj minimizes the distance, its solution is

$$\pi(\cdot|s) = \mathrm{proj}_\Delta\big(Q_h^{k,\beta,\eta}(s,\cdot)/\eta\big).$$

This finishes the proof. □

C Proofs of the Results in Tabular Setting

C.1 Proof of the Dynamic Programming Principle

Proof of Proposition 3.1. We prove a stronger version of the proposition. Specifically, there exists transition kernels $\{\widetilde{P}_t\}_{t=1}^H$, such that for all $(h,s) \in [H] \times \mathcal{S}$,

$$V_h^{\pi,\beta,\eta}(s) = \mathbb{E}_{\{\widetilde{P}_t\}_{t=h}^H}\left[\sum_{t=h}^H \Big(r_t(s_t,a_t) - \Omega_{s_t}^\eta(\pi_t) + \beta \cdot \mathrm{D}\big(P_t(\cdot|s_t,a_t) \| P_t^o(\cdot|s_t,a_t)\big)\Big) \Big| s_h = s, \pi_{h:H}\right]. \tag{C.1}$$

We prove this statement by induction. For the base case $h = H$, the claim holds trivially since no transition kernels are involved. For the inductive step, assume the conclusion holds for step $h+1$, meaning that there exists $\{\widetilde{r}_t\}_{t=1}^H$ and $\{\widetilde{P}_t\}_{t=h+1}^H$ such that

$$V_{h+1}^{\pi,\beta,\eta}(s) = \mathbb{E}_{\{\widetilde{P}_t\}_{t=h+1}^H}\left[\sum_{t=h+1}^H \Big(r_t(s_t,a_t) - \Omega_{s_t}^\eta(\pi_t) + \beta \cdot \mathrm{D}\big(P_t(\cdot|s_t,a_t) \| P_t^o(\cdot|s_t,a_t)\big)\Big) \Big| s_{h+1} = s, \pi_{h+1:H}\right]. \tag{C.2}$$

We now prove (C.1) for the case of h. For $Q_h^{\pi,\beta,\eta}(s,a)$, on the one hand, we upper bound it as

$$\begin{aligned}
&Q_h^{\pi,\beta,\eta}(s,a) \\
&= \inf_{P_t \in \Delta(S)} \mathbb{E}_{\{P_t\}_{t=h}^H}\left[\sum_{t=h}^H \Big(r_t(s_t,a_t) + \beta \cdot \mathrm{D}\big(P_t(\cdot|s_t,a_t)\big\|P_t^o(\cdot|s_t,a_t)\big)\Big) - \sum_{t=h+1}^H \Omega_{s_t}^\eta(\pi_t) \Big| s_h = s, a_h = a, \pi_{h+1:H}\right] \\
&= r_h(s,a) + \inf_{P_t \in \Delta(S)} \Bigg\{\beta \cdot \mathrm{D}\big(P_h(\cdot|s_h,a_h)\big\|P_h^o(\cdot|s_h,a_h)\big) + \int_{\mathcal{S}} P_h(\mathrm{d}s'|s,a) \mathbb{E}_{\{P_t\}_{t=h+1}^H}\Bigg[\\
&\qquad \sum_{t=h+1}^H \Big(r_t(s_t,a_t) - \Omega_{s_t}^\eta(\pi_t) + \beta \cdot \mathrm{D}\big(P_t(\cdot|s_t,a_t)\big\|P_t^o(\cdot|s_t,a_t)\big)\Big) \Big| s_{h+1} = s', \pi_{h+1:H}\Bigg]\Bigg\} \\
&\le r_h(s,a) + \inf_{P_h \in \Delta(S)} \Bigg\{\beta \cdot \mathrm{D}\big(P_h(\cdot|s_h,a_h)\big\|P_h^o(\cdot|s_h,a_h)\big) + \int_{\mathcal{S}} P_h(\mathrm{d}s'|s,a) \mathbb{E}_{\{\widetilde{P}_t\}_{t=h+1}^H}\Bigg[\\
&\qquad \sum_{t=h+1}^H \Big(r_t(s_t,a_t) - \Omega_{s_t}^\eta(\pi_t) + \beta \cdot \mathrm{D}\big(\widetilde{P}_t(\cdot|s_t,a_t)\big\|P_t^o(\cdot|s_t,a_t)\big)\Big) \Big| s_{h+1} = s', \pi_{h+1:H}\Bigg]\Bigg\} \qquad \text{(C.3)} \\
&= r_h(s,a) + \inf_{P_h \in \Delta(S)} \Big\{\beta \cdot \mathrm{D}\big(P_h(\cdot|s_h,a_h)\big\|P_h^o(\cdot|s_h,a_h)\big) + \mathbb{E}_{s' \sim P_h^o(\cdot|s,a)}\big[V_{h+1}^{\pi,\beta,\eta}(s')\big]\Big\}, \qquad \text{(C.4)}
\end{aligned}$$

where (C.3) is because the definition of the infimum operator and (C.4) is from (C.2).

On the other hand, we can lower bound $Q_h^{\pi,\beta,\eta}(s,a)$ as

$$\begin{aligned}
&Q_h^{\pi,\beta,\eta}(s,a) \\
&= \inf_{P_t \in \Delta(S)} \mathbb{E}_{\{P_t\}_{t=h}^H}\left[\sum_{t=h}^H \Big(r_t(s_t,a_t) + \beta \cdot \mathrm{D}\big(P_t(\cdot|s_t,a_t)\big\|P_t^o(\cdot|s_t,a_t)\big)\Big) - \sum_{t=h+1}^H \Omega_{s_t}^\eta(\pi_t) \Big| s_h = s, a_h = a, \pi_{h+1:H}\right] \\
&= r_h(s,a) + \inf_{P_t \in \Delta(S)} \Bigg\{\beta \cdot \mathrm{D}\big(P_h(\cdot|s_h,a_h)\big\|P_h^o(\cdot|s_h,a_h)\big) + \int_{\mathcal{S}} P_h(\mathrm{d}s'|s,a) \mathbb{E}_{\{P_t\}_{t=h+1}^H}\Bigg[\\
&\qquad \sum_{t=h+1}^H \Big(r_t(s_t,a_t) - \Omega_{s_t}^\eta(\pi_t) + \beta \cdot \mathrm{D}\big(P_t(\cdot|s_t,a_t)\big\|P_t^o(\cdot|s_t,a_t)\big)\Big) \Big| s_{h+1} = s', \pi_{h+1:H}\Bigg]\Bigg\}
\end{aligned}$$

$$\geq r_h(s,a) + \inf_{P_h\in\Delta(S)}\Bigg\{\beta\cdot \mathrm{D}\big(P_h(\cdot|s_h,a_h)\big\|P_h^o(\cdot|s_h,a_h)\big) + \int_{\mathcal{S}} P_h(\mathrm{d}s'|s,a)\inf_{P_t\in\Delta(S)}\mathbb{E}_{\{P_t\}_{t=h+1}^H}\Bigg[$$
$$\sum_{t=h+1}^H\Big(r_t(s_t,a_t) - \Omega_{s_t}^\eta(\pi_t) + \beta\cdot\mathrm{D}\big(P_t(\cdot|s_t,a_t)\big\|P_t^o(\cdot|s_t,a_t)\big)\Big)\Big|s_{h+1}=s',\pi_{h+1:H}\Bigg]\Bigg\} \tag{C.5}$$
$$= r_h(s,a) + \inf_{P_h\in\Delta(S)}\Big\{\beta\cdot\mathrm{D}\big(P_h(\cdot|s_h,a_h)\big\|P_h^o(\cdot|s_h,a_h)\big) + \mathbb{E}_{s'\sim P_h^o(\cdot|s,a)}\big[V_{h+1}^{\pi,\beta,\eta}(s')\big]\Big\}, \tag{C.6}$$

where (C.5) is because $\inf\int f\geq\int\inf f$ and (C.6) is from the definition of $V_{h+1}^{\pi,\beta,\eta}(s')$.

By combining (C.4) and (C.6), we have

$$Q_h^{\pi,\beta,\eta}(s,a) = r_h(s,a) + \inf_{\substack{\boldsymbol{\mu}_h\in\Delta(S)^d\\ P_h=\langle\boldsymbol{\phi},\boldsymbol{\mu}_h\rangle}}\Big\{\mathbb{E}_{s'\sim P_h^o(\cdot|s,a)}\big[V_{h+1}^{\pi,\beta,\eta}(s')\big] + \beta\cdot\mathrm{D}\big(P_h(\cdot|s_h,a_h)\big\|P_h^o(\cdot|s_h,a_h)\big)\Big\},$$

which implies (3.1).

Since $\Delta(\mathcal{S})$ is compact, there exists $\widetilde{P}_h$, such that

$$Q_h^{\pi,\beta,\eta}(s,a) = r_h(s,a) + \mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s,a)}\big[V_{h+1}^{\pi,\beta,\eta}(s')\big] + \beta\cdot\mathrm{D}\big(\widetilde{P}_h(\cdot|s,a)\big\|P_h^o(\cdot|s,a)\big). \tag{C.7}$$

This finishes the proof of (C.1).

For $V_h^{\pi,\beta,\eta}(s)$, on the one hand, we upper bound it as

$$V_h^{\pi,\beta,\eta}(s)$$
$$= \inf_{P_t\in\Delta(S)}\mathbb{E}_{\{P_t\}_{t=h}^H}\Bigg[\sum_{t=h}^H\Big(r_t(s_t,a_t) - \Omega_{s_t}^\eta(\pi_t) + \beta\cdot\mathrm{D}\big(P_t(\cdot|s_t,a_t)\big\|P_t^o(\cdot|s_t,a_t)\big)\Big)\Big|s_h=s,\pi_{h:H}\Bigg]$$
$$= \inf_{P_t\in\Delta(S)}\Bigg\{-\Omega_s^\eta(\pi_h) + \sum_{a\in\mathcal{A}}\pi_h(a|s)\mathbb{E}_{\{P_t\}_{t=h}^H}\Bigg[$$
$$\sum_{t=h}^H\Big(r_t(s_t,a_t) + \beta\cdot\mathrm{D}\big(P_t(\cdot|s_t,a_t)\big\|P_t^o(\cdot|s_t,a_t)\big)\Big) - \sum_{t=h+1}^H\Omega_{s_t}^\eta(\pi_t)\Big|s_h=s,a_h=a,\pi_{h+1:H}\Bigg]\Bigg\}$$
$$\leq -\Omega_s^\eta(\pi_h) + \sum_{a\in\mathcal{A}}\pi_h(a|s)\mathbb{E}_{\{\widetilde{P}_t\}_{t=h}^H}\Bigg[$$
$$\sum_{t=h}^H\Big(r_t(s_t,a_t) + \beta\cdot\mathrm{D}\big(\widetilde{P}_t(\cdot|s_t,a_t)\big\|P_t^o(\cdot|s_t,a_t)\big)\Big) - \sum_{t=h+1}^H\Omega_{s_t}^\eta(\pi_t)\Big|s_h=s,a_h=a,\pi_{h+1:H}\Bigg] \tag{C.8}$$
$$= -\Omega_s^\eta(\pi_h) + \sum_{a\in\mathcal{A}}\pi_h(a|s)Q_h^{\pi,\beta,\eta}(s,a), \tag{C.9}$$

where (C.8) is because the definition of the infimum operator and (C.9) is from (C.7) and (C.2).

On the other hand, we can lower bound $V_h^{\pi,\beta,\eta}(s)$ as

$$V_h^{\pi,\beta,\eta}(s)$$
$$= \inf_{P_t\in\Delta(S)}\mathbb{E}_{\{P_t\}_{t=h}^H}\Bigg[\sum_{t=h}^H\Big(r_t(s_t,a_t) - \Omega_{s_t}^\eta(\pi_t) + \beta\cdot\mathrm{D}\big(P_t(\cdot|s_t,a_t)\big\|P_t^o(\cdot|s_t,a_t)\big)\Big)\Big|s_h=s,\pi_{h:H}\Bigg]$$
$$= \inf_{P_t\in\Delta(S)}\Bigg\{-\Omega_s^\eta(\pi_h) + \sum_{a\in\mathcal{A}}\pi_h(a|s)\mathbb{E}_{\{P_t\}_{t=h}^H}\Bigg[$$

$$\sum_{t=h}^{H}\Big(r_t(s_t,a_t)+\beta\cdot \mathrm{D}\big(P_t(\cdot|s_t,a_t)\big\|P_t^o(\cdot|s_t,a_t)\big)\Big)-\sum_{t=h+1}^{H}\Omega_{s_t}^{\eta}(\pi_t)\Big|s_h=s,a_h=a,\pi_{h+1:H}\Big]\Big\}$$

$$\geq -\Omega_s^{\eta}(\pi_h)+\sum_{a\in\mathcal{A}}\pi_h(a|s)\inf_{P_t\in\Delta(S)}\mathbb{E}_{\{P_t\}_{t=h}^{H}}\Big[$$

$$\sum_{t=h}^{H}\Big(r_t(s_t,a_t)+\beta\cdot \mathrm{D}\big(P_t(\cdot|s_t,a_t)\big\|P_t^o(\cdot|s_t,a_t)\big)\Big)-\sum_{t=h+1}^{H}\Omega_{s_t}^{\eta}(\pi_t)\Big|s_h=s,a_h=a,\pi_{h+1:H}\Big] \tag{C.10}$$

$$=-\Omega_s^{\eta}(\pi_h)+\sum_{a\in\mathcal{A}}\pi_h(a|s)Q_h^{\pi,\beta,\eta}(s,a), \tag{C.11}$$

where (C.10) is because $\inf\sum f\geq\sum\inf f$ and (C.11) is from the definition of $Q_h^{\pi,\beta,\eta}(s,a)$.

By combining (C.9) and (C.11), we have

$$V_h^{\pi,\beta,\eta}(s)=-\Omega_s^{\eta}(\pi_h)+\sum_{a\in\mathcal{A}}\pi_h(a|s)Q_h^{\pi,\beta,\eta}(s,a),$$

which implies (3.2). □

C.2 Proof of the Double Robustness

Proof of Proposition 3.2. We have

$$\inf_{\substack{\widetilde{r}\in r+\mathcal{R}^{\eta}(\pi)\\P\in\Delta(\mathcal{S})}}\mathbb{E}_{a\sim\pi(\cdot|s)}\big[\widetilde{r}(s,a)+\mathbb{E}_{s'\sim P(\cdot|s,a)}[V(s')]+\beta\cdot \mathrm{D}\big(P(s,a)\big\|P^o(s,a)\big)\big]$$

$$=\inf_{\substack{\widetilde{r}\in r+\mathcal{R}^{\eta}(\pi)\\P\in\Delta(\mathcal{S})}}\Big(\mathbb{E}_{a\sim\pi(\cdot|s)}\big[\widetilde{r}(s,a)\big]+\mathbb{E}_{a\sim\pi(\cdot|s)}\Big[\mathbb{E}_{s'\sim P(\cdot|s,a)}[V(s')]+\beta\cdot \mathrm{D}\big(P(s,a)\big\|P^o(s,a)\big)\Big]\Big)$$

$$=\inf_{\widetilde{r}\in r+\mathcal{R}^{\eta}(\pi)}\sum_{a\in\mathcal{A}}\pi(a|s)\widetilde{r}(s,a)+\inf_{P\in\Delta(\mathcal{S})}\mathbb{E}_{a\sim\pi(\cdot|s)}\Big[\mathbb{E}_{s'\sim P(\cdot|s,a)}[V(s')]+\beta\cdot \mathrm{D}\big(P(s,a)\big\|P^o(s,a)\big)\Big]$$

$$=\inf_{\Delta r\in\mathcal{R}^{\eta}(\pi)}\sum_{a\in\mathcal{A}}\pi(a|s)\Delta r(s,a)+\inf_{P\in\Delta(\mathcal{S})}\mathbb{E}_{a\sim\pi(\cdot|s)}\Big[r(s,a)+\mathbb{E}_{s'\sim P(\cdot|s,a)}[V(s')]+\beta\cdot \mathrm{D}\big(P(s,a)\big\|P^o(s,a)\big)\Big]$$

$$=\inf_{\Delta r\in\mathcal{R}^{\eta}(\pi)}\big\langle\pi(\cdot|s),\Delta r(s,\cdot)\big\rangle+\inf_{P\in\Delta(\mathcal{S})}\mathbb{E}_{a\sim\pi(\cdot|s)}\Big[r(s,a)+\mathbb{E}_{s'\sim P(\cdot|s,a)}[V(s')]+\beta\cdot \mathrm{D}\big(P(s,a)\big\|P^o(s,a)\big)\Big]$$

$$=-\sup_{\Delta r\in\mathcal{R}^{\eta}(\pi)}\big\langle-\pi(\cdot|s),\Delta r(s,\cdot)\big\rangle+\inf_{P\in\Delta(\mathcal{S})}\mathbb{E}_{a\sim\pi(\cdot|s)}\Big[r(s,a)+\mathbb{E}_{s'\sim P(\cdot|s,a)}[V(s')]+\beta\cdot \mathrm{D}\big(P(s,a)\big\|P^o(s,a)\big)\Big]$$

$$=-\sigma_{\mathcal{R}_s^{\eta}(\pi)}\big(-\pi(\cdot|s)\big)+\inf_{P\in\Delta(\mathcal{S})}\mathbb{E}_{a\sim\pi(\cdot|s)}\Big[r(s,a)+\mathbb{E}_{s'\sim P(\cdot|s,a)}[V(s')]+\beta\cdot \mathrm{D}\big(P(s,a)\big\|P^o(s,a)\big)\Big] \tag{C.12}$$

$$=\inf_{P\in\Delta(\mathcal{S})}\mathbb{E}_{a\sim\pi(\cdot|s)}\Big[r(s,a)+\mathbb{E}_{s'\sim P(\cdot|s,a)}[V(s')]+\beta\cdot \mathrm{D}\big(P(s,a)\big\|P^o(s,a)\big)\Big]-\Omega_s^{\eta}(\pi) \tag{C.13}$$

$$=\inf_{P\in\Delta(\mathcal{S})}\mathbb{E}_{a\sim\pi(\cdot|s)}\Big[r(s,a)+\mathbb{E}_{s'\sim P(\cdot|s,a)}[V(s')]+\beta\cdot \mathrm{D}\big(P(s,a)\big\|P^o(s,a)\big)-\Omega_s^{\eta}(\pi)\Big],$$

where (C.12) is from the definition that $\sigma_{\mathcal{Z}}(\boldsymbol{y})=\max_{\boldsymbol{a}\in\mathcal{Z}}\langle\boldsymbol{a},\boldsymbol{y}\rangle$, (C.13) is because we choose $\Omega_s^{\eta}(\pi)=\sigma_{\mathcal{R}_s^{\eta}(\pi)}(-\pi(\cdot|s))$. This finishes the proof. □

C.3 Proofs of the Main Results

To establish Theorem 5.4, we first introduce Lemma C.1, which bounds the regret by the sum of bonus terms under the expectation defined by the worst-case transition.

Lemma C.1. Under Assumption 5.3, for each f-divergence setting and any $\delta \in (0, 1/2)$, with probability at least $1 - 2\delta$, the regret of Algorithm 1 satisfies

$$\mathrm{Regret}(K) \le 2 \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{P^{w,k}, \pi^k} \left[b_h^k(s, a) \right], \tag{C.14}$$

where $b_h^k(s, a)$ is the bonus term defined in Section 4.2 for each f-divergence setting.

To upper bound the sum of the expectations of $\sqrt{\frac{1}{n_h^{k-1} \vee 1}}$ under the worst-case environment, He et al. (2025) proved the following lemma.

Lemma C.2. (He et al., 2025, Lemma C.4, Lemma C.9) If the algorithm RSPVI selects π^k for the k-th episode, with probability at least $1 - \delta$, it holds that

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{P^{w,k}, \pi^k} \left[\sqrt{\frac{1}{n_h^{k-1}(s, a) \vee 1}} \right] = \widetilde{\mathcal{O}}\big(\sqrt{C_{vr} S A H^2 K} + C_{vr} S A H \big).$$

Proof of Theorem 5.4. The result follows immediately from combining Lemma C.1 and Lemma C.2, and substituting the specific bonus formulation defined in Section 4.2. □

C.4 Proofs of the Technical Lemmas

Here, we present the proof of Lemma C.1 is various f-divergence settings. For convenience, we also write $P^{w,k} := P^{w,\pi^k}$, $\mathrm{d}^k := \mathrm{d}^{\pi^k}$ and $\mathrm{q}^k := \mathrm{q}^{\pi^k}$.

C.4.1 Proof of Lemma C.1 in TV Setting

Lemma C.3 (Dual formulation)**.** (He et al., 2025, Lemma D.1) For the optimization problem $Q_h(s, a) = r_h(s, a) + \inf_{P \in \Delta(S)} \big(\mathbb{E}_P[V_{h+1}] + \beta \mathrm{TV}(P \| P_h^o) \big)(s, a)$, we have its dual formulation as follows

$$Q_h = r_h - \mathbb{E}_{P_h^o} \Big[\Big(\min_{s \in \mathcal{S}} V_{h+1}(s) + \beta - V_{h+1}(s) \Big)_+ \Big] + \Big(\min_{s \in \mathcal{S}} V_{h+1}(s) + \beta \Big). \tag{C.15}$$

Lemma C.4 (Optimism)**.** If we set the bonus term as follows

$$\mathrm{bonus}_h^k(s, a) = 2H \sqrt{\frac{2S \ln(2SAHK/\delta)}{n_h^{k-1}(s, a) \vee 1}}, \tag{C.16}$$

then for any policy π and any $(k, h, s, a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$, with probability at least $1 - 2\delta$, we have $Q_h^{k,\beta,\eta}(s, a) \ge Q_h^{\pi,\beta,\eta}(s, a)$. Specially, by setting $\pi = \pi^*$, we have $Q_h^{k,\beta,\eta}(s, a) \ge Q_h^{\pi^*,\beta,\eta}(s, a)$.

Proof. We prove this by induction. First, when $h = H + 1$, $Q_{H+1}^{k,\beta,\eta}(s, a) = 0 = Q_{H+1}^{\pi,\beta,\eta}(s, a)$ holds trivially.

Assume $Q_{h+1}^{k,\beta,\eta}(s, a) \ge Q_{h+1}^{\pi,\beta,\eta}(s, a)$ holds, due to the choice of π_{h+1}^k in (4.3), we have

$$\begin{aligned} V_{h+1}^{k,\beta,\eta}(s) &= \big\langle Q_{h+1}^{k,\beta,\eta}(s, \cdot), \pi_{h+1}^k(\cdot|s) \big\rangle - \Omega_s^\eta(\pi_{h+1}^k) \\ &\ge \big\langle Q_{h+1}^{k,\beta,\eta}(s, \cdot), \pi_{h+1}(\cdot|s) \big\rangle - \Omega_s^\eta(\pi_{h+1}) \end{aligned}$$

$$\geq \left\langle Q_{h+1}^{\pi,\beta,\eta}(s,\cdot), \pi_{h+1}(\cdot|s) \right\rangle - \Omega_s^\eta(\pi_{h+1}) = V_{h+1}^{\pi,\beta,\eta}(s).$$

Recall that we denote $Q_h^{k,\beta,\eta}$ as the optimistic estimation in k-th episode, that is,

$$Q_h^{k,\beta,\eta}(s,a) = \min\left\{ \text{bonus}_h^k(s,a) + \widehat{r}_h^k(s,a) + \inf_{P\in\Delta(S)} \left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right] + \beta \text{TV}(P\|\widehat{P}_h^k)\right)(s,a), H-h+1 \right\}.$$

If $Q_h^{k,\beta,\eta}(s,a) = H-h+1$, then it follows immediately that

$$Q_h^{k,\beta,\eta}(s,a) = H-h+1 \geq Q_h^{\pi,\beta,\eta}(s,a)$$

by the definition of $Q_h^{\pi,\beta,\eta}(s,a)$. Otherwise, we can infer that

$$\begin{aligned}
&Q_h^{k,\beta,\eta} - Q_h^{\pi,\beta,\eta} \\
&= \text{bonus}_h^k + \widehat{r}_h^k + \inf_{P\in\Delta(S)} \left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right] + \beta \text{TV}(P\|\widehat{P}_h^k)\right) - r_h - \inf_{P\in\Delta(S)} \left(\mathbb{E}_P\left[V_{h+1}^{\pi,\beta,\eta}\right] + \beta \text{TV}(P\|P_h^o)\right) \\
&= \text{bonus}_h^k + \widehat{r}_h^k - r_h + \inf_{P\in\Delta(S)} \left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right] + \beta \text{TV}(P\|\widehat{P}_h^k)\right) - \inf_{P\in\Delta(S)} \left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right] + \beta \text{TV}(P\|P_h^o)\right) \\
&\quad + \inf_{P\in\Delta(S)} \left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right] + \beta \text{TV}(P\|P_h^o)\right) - \inf_{P\in\Delta(S)} \left(\mathbb{E}_P\left[V_{h+1}^{\pi,\beta,\eta}\right] + \beta \text{TV}(P\|P_h^o)\right) \\
&\geq \text{bonus}_h^k + \widehat{r}_h^k - r_h + \inf_{P\in\Delta(S)} \left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right] - \mathbb{E}_P\left[V_{h+1}^{\pi,\beta,\eta}\right]\right) \\
&\quad + \inf_{P\in\Delta(S)} \left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right] + \beta \text{TV}(P\|\widehat{P}_h^k)\right) - \inf_{P\in\Delta(S)} \left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right] + \beta \text{TV}(P\|P_h^o)\right) \qquad \text{(C.17)} \\
&\geq \text{bonus}_h^k + \widehat{r}_h^k - r_h + \inf_{P\in\Delta(S)} \left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right] + \beta \text{TV}(P\|\widehat{P}_h^k)\right) - \inf_{P\in\Delta(S)} \left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right] + \beta \text{TV}(P\|P_h^o)\right) \qquad \text{(C.18)} \\
&= \text{bonus}_h^k + \widehat{r}_h^k - r_h - \mathbb{E}_{P_h^o}\left[\left(\min_{s\in\mathcal{S}} V_{h+1}^{k,\beta,\eta}(s) + \beta - V_{h+1}^{k,\beta,\eta}(s)\right)_+\right] + \left(\min_{s\in\mathcal{S}} V_{h+1}^{k,\beta,\eta}(s) + \beta\right) \\
&\quad + \mathbb{E}_{\widehat{P}_h^k}\left[\left(\min_{s\in\mathcal{S}} V_{h+1}^{k,\beta,\eta}(s) + \beta - V_{h+1}^{k,\beta,\eta}(s)\right)_+\right] - \left(\min_{s\in\mathcal{S}} V_{h+1}^{k,\beta,\eta}(s) + \beta\right) \qquad \text{(C.19)} \\
&= \text{bonus}_h^k + \widehat{r}_h^k - r_h \\
&\quad - \mathbb{E}_{P_h^o}\left[\left(\min_{s\in\mathcal{S}} V_{h+1}^{k,\beta,\eta}(s) + \beta - V_{h+1}^{k,\beta,\eta}(s)\right)_+\right] + \mathbb{E}_{\widehat{P}_h^k}\left[\left(\min_{s\in\mathcal{S}} V_{h+1}^{k,\beta,\eta}(s) + \beta - V_{h+1}^{k,\beta,\eta}(s)\right)_+\right] \\
&\geq \text{bonus}_h^k - \underbrace{\left|\widehat{r}_h^k - r_h\right|}_{\text{(i)}} \\
&\quad - \underbrace{\left|\mathbb{E}_{P_h^o}\left[\left(\min_{s\in\mathcal{S}} V_{h+1}^{k,\beta,\eta}(s) + \beta - V_{h+1}^{k,\beta,\eta}(s)\right)_+\right] - \mathbb{E}_{\widehat{P}_h^k}\left[\left(\min_{s\in\mathcal{S}} V_{h+1}^{k,\beta,\eta}(s) + \beta - V_{h+1}^{k,\beta,\eta}(s)\right)_+\right]\right|}_{\text{(ii)}}, \qquad \text{(C.20)}
\end{aligned}$$

where (C.17) is from $\inf f(x) - \inf g(x) \geq \inf(f-g)(x)$, (C.18) is from the induction assumption, we plug in the dual formulation (C.15) in (C.19).

For term (i) in (C.20), from Lemma E.1 and a union bound, with probability at least $1-\delta$, we have

$$\left|\widehat{r}_h^k(s,a) - r_h(s,a)\right| \leq \sqrt{\frac{\ln(2SAHK/\delta)}{2n_h^{k-1}(s,a) \vee 1}} \qquad \text{(C.21)}$$

for any $(k,h,s,a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$.

For any fixed V, we apply Lemma E.3 and have

$$\left|\mathbb{E}_{P_h^o}[V] - \mathbb{E}_{\widehat{P}_h^k}[V]\right| \le \left\|P_h^o - \widehat{P}_h^k\right\|_1 \cdot \|V\|_\infty \le H\sqrt{\frac{2S\ln(2/\delta)}{n_h^{k-1} \vee 1}}, \tag{C.22}$$

with probability at least $1-\delta$.

For term (ii) in (C.20), by applying (C.22) and a union bound, with probability at least $1-\delta$, we have

$$\begin{aligned}&\left|\mathbb{E}_{P_h^o}\left[\left(\min_{s\in\mathcal{S}} V_{h+1}^{k,\beta,\eta}(s) + \beta - V_{h+1}^{k,\beta,\eta}(s)\right)_+\right] - \mathbb{E}_{\widehat{P}_h^k}\left[\left(\min_{s\in\mathcal{S}} V_{h+1}^{k,\beta,\eta}(s) + \beta - V_{h+1}^{k,\beta,\eta}(s)\right)_+\right]\right| \\ &\le H\sqrt{\frac{2S\ln(2SAHK/\delta)}{n_h^{k-1} \vee 1}}\end{aligned} \tag{C.23}$$

for any $(k,h,s,a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$.

Apply the union bound again and combine (C.20) with (C.21), (C.23), the definition of bonus and induction assumption. With probability at least $1-2\delta$, we have $Q_h^{k,\rho}(s,a) \ge Q_h^{\pi,\rho}(s,a)$ for any $(k,h,s,a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$. This completes the proof. □

Lemma C.5. If we set the bonus term to be the same as in Lemma C.4, then for any $(k,s,a) \in [K] \times \mathcal{S} \times \mathcal{A}$, with probability at least $1-\delta$, the sum of estimation errors can be bounded as

$$Q_1^{k,\beta,\eta}(s,a) - Q_1^{\pi^k,\beta,\eta}(s,a) \le 2 \cdot \mathbb{E}_{P^{w,k},\pi^k}\left[\text{bonus}_h^k\right],$$

Proof. From the proof of Lemma C.4, we see that with probability at least $1-\delta$, for any $(k,s,a) \in [K] \times \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}&\left|\widehat{r}_h^k(s,a) - r_h(s,a)\right| + \Big|\inf_{P\in\Delta(S)}\left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right] + \beta\text{TV}(P\|\widehat{P}_h^k)\right)(s,a) \\ &\qquad - \inf_{P\in\Delta(S)}\left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right] + \beta\text{TV}(P\|P_h^o)\right)(s,a)\Big| \le \text{bonus}_h^k(s,a).\end{aligned} \tag{C.24}$$

Recall that we define $P_h^{w,k} = \operatorname{argmin}_{P\in\Delta(S)}\left(\mathbb{E}_P\left[V_{h+1}^{\pi^k,\beta,\eta}\right] + \beta\text{TV}(P\|P_h^o)\right)$ as the worst-case transition in Definition 5.1, we have

$$\begin{aligned}&Q_h^{k,\beta,\eta} - Q_h^{\pi^k,\beta,\eta} \\ &\le \text{bonus}_h^k + \widehat{r}_h^k + \inf_{P\in\Delta(S)}\left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right] + \beta\text{TV}(P\|\widehat{P}_h^k)\right) - r_h - \inf_{P\in\Delta(S)}\left(\mathbb{E}_P\left[V_{h+1}^{\pi^k,\beta,\eta}\right] + \beta\text{TV}(P\|P_h^o)\right) \\ &\le 2\text{bonus}_h^k + \inf_{P\in\Delta(S)}\left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right] + \beta\text{TV}(P\|P_h^o)\right) - \inf_{P\in\Delta(S)}\left(\mathbb{E}_P\left[V_{h+1}^{\pi^k,\beta,\eta}\right] + \beta\text{TV}(P\|P_h^o)\right) \qquad \text{(C.25)} \\ &= 2\text{bonus}_h^k + \mathbb{E}_{P_h^{w,k},\pi_h^k}\left[Q_{h+1}^{k,\beta,\eta} - Q_{h+1}^{\pi^k,\beta,\eta}\right]. \qquad \text{(C.26)}\end{aligned}$$

where (C.25) uses (C.24). Apply (C.26) recursively, we can obtain the result. □

Proof of Lemma C.1 in TV Setting. The result directly follows from combining Lemma C.4 and Lemma C.5, along with applying a union bound. □

C.4.2 Proof of Lemma C.1 in KL Setting

Lemma C.6 (Dual formulation)**.** (He et al., 2025, Lemma D.5) For the optimization problem $Q_h(s,a) = r_h(s,a) + \inf_{P\in\Delta(S)} \big(\mathbb{E}_P[V_{h+1}] + \beta \mathrm{KL}(P\|P_h^o)\big)(s,a)$, we have its dual formulation as follows

$$Q_h = r_h - \beta \ln \mathbb{E}_{P_h^o}\big[e^{-\beta^{-1}V_{h+1}}\big]. \tag{C.27}$$

Lemma C.7 (Optimism)**.** If we set the bonus term as follows

$$\mathrm{bonus}_h^k(s,a) = \big(1+\beta e^{\beta^{-1}H}\sqrt{S}\big)\sqrt{\frac{2\ln(2SAHK/\delta)}{n_h^{k-1}(s,a)\vee 1}}, \tag{C.28}$$

then for any policy π and any $(k,h,s,a)\in[K]\times[H]\times\mathcal{S}\times\mathcal{A}$, with probability at least $1-2\delta$, we have $Q_h^{k,\beta,\eta}(s,a) \geq Q_h^{\pi,\beta,\eta}(s,a)$. Specially, by setting $\pi=\pi^*$, we have $Q_h^{k,\beta,\eta}(s,a) \geq Q_h^{\pi^*,\beta,\eta}(s,a)$.

Proof. We prove this by induction. First, when $h=H+1$, $Q_{H+1}^{k,\beta,\eta}(s,a) = 0 = Q_{H+1}^{\pi,\beta,\eta}(s,a)$ holds trivially.

Assume $Q_{h+1}^{k,\beta,\eta}(s,a) \geq Q_{h+1}^{\pi,\beta,\eta}(s,a)$ holds, due to the choice of π_{h+1}^k in (4.3), we have

$$\begin{aligned}
V_{h+1}^{k,\beta,\eta}(s) &= \big\langle Q_{h+1}^{k,\beta,\eta}(s,\cdot), \pi_{h+1}^k(\cdot|s)\big\rangle - \Omega_s^\eta(\pi_{h+1}^k)\\
&\geq \big\langle Q_{h+1}^{k,\beta,\eta}(s,\cdot), \pi_{h+1}(\cdot|s)\big\rangle - \Omega_s^\eta(\pi_{h+1})\\
&\geq \big\langle Q_{h+1}^{\pi,\beta,\eta}(s,\cdot), \pi_{h+1}(\cdot|s)\big\rangle - \Omega_s^\eta(\pi_{h+1}) = V_{h+1}^{\pi,\beta,\eta}(s).
\end{aligned}$$

Recall that we denote $Q_h^{k,\beta,\eta}$ as the optimistic estimation in k-th episode, that is,

$$Q_h^{k,\beta,\eta}(s,a) = \min\Big\{\mathrm{bonus}_h^k(s,a) + \widehat{r}_h^k(s,a) + \inf_{P\in\Delta(S)}\big(\mathbb{E}_P\big[V_{h+1}^{k,\beta,\eta}\big] + \beta\mathrm{KL}(P\|\widehat{P}_h^k)\big)(s,a), H-h+1\Big\}.$$

If $Q_h^{k,\beta,\eta}(s,a) = H-h+1$, then it follows immediately that

$$Q_h^{k,\beta,\eta}(s,a) = H-h+1 \geq Q_h^{\pi,\beta,\eta}(s,a)$$

by the definition of $Q_h^{\pi,\beta,\eta}(s,a)$. Otherwise, we can infer that

$$\begin{aligned}
&Q_h^{k,\beta,\eta} - Q_h^{\pi,\beta,\eta}\\
&= \mathrm{bonus}_h^k + \widehat{r}_h^k + \inf_{P\in\Delta(S)}\big(\mathbb{E}_P\big[V_{h+1}^{k,\beta,\eta}\big] + \beta\mathrm{KL}(P\|\widehat{P}_h^k)\big) - r_h - \inf_{P\in\Delta(S)}\big(\mathbb{E}_P\big[V_{h+1}^{\pi,\beta,\eta}\big] + \beta\mathrm{KL}(P\|P_h^o)\big)\\
&= \mathrm{bonus}_h^k + \widehat{r}_h^k - r_h + \inf_{P\in\Delta(S)}\big(\mathbb{E}_P\big[V_{h+1}^{k,\beta,\eta}\big] + \beta\mathrm{KL}(P\|\widehat{P}_h^k)\big) - \inf_{P\in\Delta(S)}\big(\mathbb{E}_P\big[V_{h+1}^{k,\beta,\eta}\big] + \beta\mathrm{KL}(P\|P_h^o)\big)\\
&\quad + \inf_{P\in\Delta(S)}\big(\mathbb{E}_P\big[V_{h+1}^{k,\beta,\eta}\big] + \beta\mathrm{KL}(P\|P_h^o)\big) - \inf_{P\in\Delta(S)}\big(\mathbb{E}_P\big[V_{h+1}^{\pi,\beta,\eta}\big] + \beta\mathrm{KL}(P\|P_h^o)\big)\\
&\geq \mathrm{bonus}_h^k + \widehat{r}_h^k - r_h + \inf_{P\in\Delta(S)}\big(\mathbb{E}_P\big[V_{h+1}^{k,\beta,\eta}\big] - \mathbb{E}_P\big[V_{h+1}^{\pi,\beta,\eta}\big]\big)\\
&\quad + \inf_{P\in\Delta(S)}\big(\mathbb{E}_P\big[V_{h+1}^{k,\beta,\eta}\big] + \beta\mathrm{KL}(P\|\widehat{P}_h^k)\big) - \inf_{P\in\Delta(S)}\big(\mathbb{E}_P\big[V_{h+1}^{k,\beta,\eta}\big] + \beta\mathrm{KL}(P\|P_h^o)\big) \qquad \text{(C.29)}\\
&\geq \mathrm{bonus}_h^k + \widehat{r}_h^k - r_h + \inf_{P\in\Delta(S)}\big(\mathbb{E}_P\big[V_{h+1}^{k,\beta,\eta}\big] + \beta\mathrm{KL}(P\|\widehat{P}_h^k)\big) - \inf_{P\in\Delta(S)}\big(\mathbb{E}_P\big[V_{h+1}^{k,\beta,\eta}\big] + \beta\mathrm{KL}(P\|P_h^o)\big) \qquad \text{(C.30)}
\end{aligned}$$

$$= \text{bonus}_h^k + \widehat{r}_h^k - r_h + \beta \ln \mathbb{E}_{P_h^o}\left[e^{-\beta^{-1}V_{h+1}^{k,\beta,\eta}}\right] - \beta \ln \mathbb{E}_{\widehat{P}_h^k}\left[e^{-\beta^{-1}V_{h+1}^{k,\beta,\eta}}\right] \tag{C.31}$$

$$\geq \text{bonus}_h^k - \underbrace{\left|\widehat{r}_h^k - r_h\right|}_{\text{(i)}} - \underbrace{\beta\left|\ln \mathbb{E}_{\widehat{P}_h^k}\left[e^{-\beta^{-1}V_{h+1}^{k,\beta,\eta}}\right] - \ln \mathbb{E}_{P_h^o}\left[e^{-\beta^{-1}V_{h+1}^{k,\beta,\eta}}\right]\right|}_{\text{(ii)}}, \tag{C.32}$$

where (C.29) is from $\inf f(x) - \inf g(x) \geq \inf(f-g)(x)$, (C.30) is from the induction assumption, we plug in the dual formulation (C.27) in (C.31).

For term (i) in (C.32), from Lemma E.1 and a union bound, with probability at least $1-\delta$, we have

$$\left|\widehat{r}_h^k(s,a) - r_h(s,a)\right| \leq \sqrt{\frac{\ln(2SAHK/\delta)}{2n_h^{k-1}(s,a) \vee 1}} \tag{C.33}$$

for any $(k,h,s,a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$.

For term (ii) in (C.32), by applying (C.22) and a union bound, with probability at least $1-\delta$, we have

$$\left|\ln \mathbb{E}_{\widehat{P}_h^k}\left[e^{-\beta^{-1}V_{h+1}^{k,\beta,\eta}}\right] - \ln \mathbb{E}_{P_h^o}\left[e^{-\beta^{-1}V_{h+1}^{k,\beta,\eta}}\right]\right| \leq e^{\beta^{-1}H}\sqrt{\frac{2S\ln(2SAHK/\delta)}{n_h^{k-1} \vee 1}} \tag{C.34}$$

for any $(k,h,s,a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$.

Apply the union bound again and combine (C.32) with (C.33), (C.34), the definition of bonus and induction assumption. With probability at least $1-2\delta$, we have $Q_h^{k,\rho}(s,a) \geq Q_h^{\pi,\rho}(s,a)$ for any $(k,h,s,a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$. This completes the proof. □

Lemma C.8. If we set the bonus term to be the same as in Lemma C.7, then for any $(k,s,a) \in [K] \times \mathcal{S} \times \mathcal{A}$, with probability at least $1-\delta$, the sum of estimation errors can be bounded as

$$Q_1^{k,\beta,\eta}(s,a) - Q_1^{\pi^k,\beta,\eta}(s,a) \leq 2 \cdot \mathbb{E}_{P^{w,k},\pi^k}\left[\text{bonus}_h^k\right],$$

Proof. From the proof of Lemma C.7, we see that with probability at least $1-\delta$, for any $(k,s,a) \in [K] \times \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}&\left|\widehat{r}_h^k(s,a) - r_h(s,a)\right| + \Big|\inf_{P \in \Delta(S)}\left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right] + \beta \text{KL}(P\|\widehat{P}_h^k)\right)(s,a) \\ &\qquad - \inf_{P \in \Delta(S)}\left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right] + \beta \text{KL}(P\|P_h^o)\right)(s,a)\Big| \leq \text{bonus}_h^k(s,a).\end{aligned} \tag{C.35}$$

Recall that we define $P_h^{w,k} = \operatorname{argmin}_{P \in \Delta(S)}\left(\mathbb{E}_P\left[V_{h+1}^{\pi^k,\beta,\eta}\right] + \beta \text{KL}(P\|P_h^o)\right)$ as the worst-case transition in Definition 5.1, we have

$$\begin{aligned}&Q_h^{k,\beta,\eta} - Q_h^{\pi^k,\beta,\eta} \\ &\leq \text{bonus}_h^k + \widehat{r}_h^k + \inf_{P \in \Delta(S)}\left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right] + \beta \text{KL}(P\|\widehat{P}_h^k)\right) - r_h - \inf_{P \in \Delta(S)}\left(\mathbb{E}_P\left[V_{h+1}^{\pi^k,\beta,\eta}\right] + \beta \text{KL}(P\|P_h^o)\right) \\ &\leq 2\text{bonus}_h^k + \inf_{P \in \Delta(S)}\left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right] + \beta \text{KL}(P\|P_h^o)\right) - \inf_{P \in \Delta(S)}\left(\mathbb{E}_P\left[V_{h+1}^{\pi^k,\beta,\eta}\right] + \beta \text{KL}(P\|P_h^o)\right) \qquad \text{(C.36)} \\ &= 2\text{bonus}_h^k + \mathbb{E}_{P_h^{w,k},\pi_h^k}\left[Q_{h+1}^{k,\beta,\eta} - Q_{h+1}^{\pi^k,\beta,\eta}\right]. \qquad \text{(C.37)}\end{aligned}$$

where (C.36) uses (C.35). Apply (C.37) recursively, we can obtain the result. □

Proof of Lemma C.1 in KL Setting. The result directly follows from combining Lemma C.7 and Lemma C.8, along with applying a union bound. □

C.4.3 Proof of Lemma C.1 in χ^2 Setting

Lemma C.9 (Dual formulation)**.** (He et al., 2025, Lemma D.9) For the optimization problem $Q_h(s,a) = r_h(s,a) + \inf_{P\in\Delta(S)} \big(\mathbb{E}_P[V_{h+1}] + \beta\chi^2(P\|P_h^o)\big)(s,a)$, we have its dual formulation as follows

$$Q_h = r_h + \sup_{\boldsymbol{\lambda}\in[0,H]} \big(\mathbb{E}_{P_h^o}\big[V_{h+1} - \boldsymbol{\lambda}\big] - \frac{1}{4\beta}\mathrm{Var}_{P_h^o}(V_{h+1} - \boldsymbol{\lambda})\big). \tag{C.38}$$

Lemma C.10 (Optimism)**.** If we set the bonus term as follows

$$\mathrm{bonus}_h^k(s,a) = \left(2 + \frac{3H}{4\beta}\right) H\sqrt{\frac{2S^2\ln(48SAH^3K^2/\delta)}{n_h^{k-1}(s,a)\vee 1}} + \left(1 + \frac{1}{4\beta}\right)\frac{1}{K}, \tag{C.39}$$

then for any policy π and any $(k,h,s,a)\in[K]\times[H]\times\mathcal{S}\times\mathcal{A}$, with probability at least $1-3\delta$, we have $Q_h^{k,\beta,\eta}(s,a) \geq Q_h^{\pi,\beta,\eta}(s,a)$. Specially, by setting $\pi = \pi^*$, we have $Q_h^{k,\beta,\eta}(s,a) \geq Q_h^{\pi^*,\beta,\eta}(s,a)$.

Proof. We prove this by induction. First, when $h = H+1$, $Q_{H+1}^{k,\beta,\eta}(s,a) = 0 = Q_{H+1}^{\pi,\beta,\eta}(s,a)$ holds trivially.

Assume $Q_{h+1}^{k,\beta,\eta}(s,a) \geq Q_{h+1}^{\pi,\beta,\eta}(s,a)$ holds, due to the choice of π_{h+1}^k in (4.3), we have

$$\begin{aligned}
V_{h+1}^{k,\beta,\eta}(s) &= \big\langle Q_{h+1}^{k,\beta,\eta}(s,\cdot), \pi_{h+1}^k(\cdot|s)\big\rangle - \Omega_s^\eta(\pi_{h+1}^k) \\
&\geq \big\langle Q_{h+1}^{k,\beta,\eta}(s,\cdot), \pi_{h+1}(\cdot|s)\big\rangle - \Omega_s^\eta(\pi_{h+1}) \\
&\geq \big\langle Q_{h+1}^{\pi,\beta,\eta}(s,\cdot), \pi_{h+1}(\cdot|s)\big\rangle - \Omega_s^\eta(\pi_{h+1}) = V_{h+1}^{\pi,\beta,\eta}(s).
\end{aligned}$$

Recall that we denote $Q_h^{k,\beta,\eta}$ as the optimistic estimation in k-th episode, that is,

$$Q_h^{k,\beta,\eta}(s,a) = \min\left\{\mathrm{bonus}_h^k(s,a) + \widehat{r}_h^k(s,a) + \inf_{P\in\Delta(S)}\big(\mathbb{E}_P\big[V_{h+1}^{k,\beta,\eta}\big] + \beta\chi^2(P\|\widehat{P}_h^k)\big)(s,a), H-h+1\right\}.$$

If $Q_h^{k,\beta,\eta}(s,a) = H-h+1$, then it follows immediately that

$$Q_h^{k,\beta,\eta}(s,a) = H-h+1 \geq Q_h^{\pi,\beta,\eta}(s,a)$$

by the definition of $Q_h^{\pi,\beta,\eta}(s,a)$. Otherwise, we can infer that

$$\begin{aligned}
&Q_h^{k,\beta,\eta} - Q_h^{\pi,\beta,\eta} \\
&= \mathrm{bonus}_h^k + \widehat{r}_h^k + \inf_{P\in\Delta(S)}\big(\mathbb{E}_P\big[V_{h+1}^{k,\beta,\eta}\big] + \beta\chi^2(P\|\widehat{P}_h^k)\big) - r_h - \inf_{P\in\Delta(S)}\big(\mathbb{E}_P\big[V_{h+1}^{\pi,\beta,\eta}\big] + \beta\chi^2(P\|P_h^o)\big) \\
&= \mathrm{bonus}_h^k + \widehat{r}_h^k - r_h + \inf_{P\in\Delta(S)}\big(\mathbb{E}_P\big[V_{h+1}^{k,\beta,\eta}\big] + \beta\chi^2(P\|\widehat{P}_h^k)\big) - \inf_{P\in\Delta(S)}\big(\mathbb{E}_P\big[V_{h+1}^{k,\beta,\eta}\big] + \beta\chi^2(P\|P_h^o)\big) \\
&\quad + \inf_{P\in\Delta(S)}\big(\mathbb{E}_P\big[V_{h+1}^{k,\beta,\eta}\big] + \beta\chi^2(P\|P_h^o)\big) - \inf_{P\in\Delta(S)}\big(\mathbb{E}_P\big[V_{h+1}^{\pi,\beta,\eta}\big] + \beta\chi^2(P\|P_h^o)\big) \\
&\geq \mathrm{bonus}_h^k + \widehat{r}_h^k - r_h + \inf_{P\in\Delta(S)}\big(\mathbb{E}_P\big[V_{h+1}^{k,\beta,\eta}\big] - \mathbb{E}_P\big[V_{h+1}^{\pi,\beta,\eta}\big]\big)
\end{aligned}$$

$$+ \inf_{P\in\Delta(S)} \left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right] + \beta\chi^2(P\|\widehat{P}_h^k)\right) - \inf_{P\in\Delta(S)} \left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right] + \beta\chi^2(P\|P_h^o)\right) \tag{C.40}$$

$$\geq \text{bonus}_h^k + \widehat{r}_h^k - r_h + \inf_{P\in\Delta(S)} \left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right] + \beta\chi^2(P\|\widehat{P}_h^k)\right) - \inf_{P\in\Delta(S)} \left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right] + \beta\chi^2(P\|P_h^o)\right) \tag{C.41}$$

$$= \text{bonus}_h^k + \widehat{r}_h^k - r_h + \sup_{\boldsymbol{\lambda}\in[0,H]} \left(\mathbb{E}_{\widehat{P}_h^k}\left[V_{h+1}^{k,\beta,\eta} - \boldsymbol{\lambda}\right] - \frac{1}{4\beta}\text{Var}_{\widehat{P}_h^k}\left(V_{h+1}^{k,\beta,\eta} - \boldsymbol{\lambda}\right)\right)$$
$$- \sup_{\boldsymbol{\lambda}\in[0,H]} \left(\mathbb{E}_{P_h^o}\left[V_{h+1}^{k,\beta,\eta} - \boldsymbol{\lambda}\right] - \frac{1}{4\beta}\text{Var}_{P_h^o}\left(V_{h+1}^{k,\beta,\eta} - \boldsymbol{\lambda}\right)\right) \tag{C.42}$$

$$\geq \text{bonus}_h^k + \widehat{r}_h^k - r_h + \inf_{\boldsymbol{\lambda}\in[0,H]} \left\{\left(\mathbb{E}_{\widehat{P}_h^k}\left[V_{h+1}^{k,\beta,\eta} - \boldsymbol{\lambda}\right] - \frac{1}{4\beta}\text{Var}_{\widehat{P}_h^k}\left(V_{h+1}^{k,\beta,\eta} - \boldsymbol{\lambda}\right)\right)\right.$$
$$\left. - \left(\mathbb{E}_{P_h^o}\left[V_{h+1}^{k,\beta,\eta} - \boldsymbol{\lambda}\right] - \frac{1}{4\beta}\text{Var}_{P_h^o}\left(V_{h+1}^{k,\beta,\eta} - \boldsymbol{\lambda}\right)\right)\right\}$$
$$\geq \text{bonus}_h^k - \left|\widehat{r}_h^k - r_h\right| - \sup_{\boldsymbol{\lambda}\in[0,H]} \left|\left(\mathbb{E}_{\widehat{P}_h^k}\left[V_{h+1}^{k,\beta,\eta} - \boldsymbol{\lambda}\right] - \frac{1}{4\beta}\text{Var}_{\widehat{P}_h^k}\left(V_{h+1}^{k,\beta,\eta} - \boldsymbol{\lambda}\right)\right)\right.$$
$$\left. - \left(\mathbb{E}_{P_h^o}\left[V_{h+1}^{k,\beta,\eta} - \boldsymbol{\lambda}\right] - \frac{1}{4\beta}\text{Var}_{P_h^o}\left(V_{h+1}^{k,\beta,\eta} - \boldsymbol{\lambda}\right)\right)\right|$$
$$\geq \text{bonus}_h^k - \underbrace{\left|\widehat{r}_h^k - r_h\right|}_{\text{(i)}} - \underbrace{\sup_{\boldsymbol{\lambda}\in[0,H]} \left|\mathbb{E}_{\widehat{P}_h^k}\left[V_{h+1}^{k,\beta,\eta} - \boldsymbol{\lambda}\right] - \mathbb{E}_{P_h^o}\left[V_{h+1}^{k,\beta,\eta} - \boldsymbol{\lambda}\right]\right|}_{\text{(ii)}}$$
$$\underbrace{- \frac{1}{4\beta} \sup_{\boldsymbol{\lambda}\in[0,H]} \left|\text{Var}_{\widehat{P}_h^k}\left(V_{h+1}^{k,\beta,\eta} - \boldsymbol{\lambda}\right) - \text{Var}_{P_h^o}\left(V_{h+1}^{k,\beta,\eta} - \boldsymbol{\lambda}\right)\right|}_{\text{(iii)}}, \tag{C.43}$$

where (C.40) is from $\inf f(x) - \inf g(x) \geq \inf(f-g)(x)$, (C.41) is from the induction assumption, we plug in the dual formulation (C.38) in (C.42).

For term (i) in (C.43), from Lemma E.1 and a union bound, with probability at least $1-\delta$, we have

$$\left|\widehat{r}_h^k(s,a) - r_h(s,a)\right| \leq \sqrt{\frac{\ln(2SAHK/\delta)}{2n_h^{k-1}(s,a) \vee 1}}, \tag{C.44}$$

for any $(k,h,s,a) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}$.

We denote $V(\boldsymbol{\lambda}) = V_{h+1}^{k,\rho} - \boldsymbol{\lambda} \in [-H,H]$ and $\mathcal{V} = \left\{V \in \mathbb{R}^S : \|V\|_\infty \leq H\right\}$. To bound term (ii) in (C.43), we create a ϵ-net $\mathcal{N}_\mathcal{V}(\epsilon)$ for $\mathcal{V}$. From Lemma E.4, it holds that $\ln|\mathcal{N}_\mathcal{V}(\epsilon)| \leq |S|\cdot\ln(3H/\epsilon)$.

Therefore, by the definition of $\mathcal{N}_\mathcal{V}(\epsilon)$, for any fixed V, there exists a $V' \in \mathcal{N}_\mathcal{V}(\epsilon)$ such that $\|V - V'\|_\infty \leq \epsilon$, that is

$$\begin{aligned}\left|\mathbb{E}_{P_h^o}[V] - \mathbb{E}_{\widehat{P}_h^k}[V]\right| &\leq \left|\mathbb{E}_{P_h^o}[V] - \mathbb{E}_{P_h^o}[V']\right| + \left|\mathbb{E}_{P_h^o}[V'] - \mathbb{E}_{\widehat{P}_h^k}[V']\right| + \left|\mathbb{E}_{\widehat{P}_h^k}[V'] - \mathbb{E}_{\widehat{P}_h^k}[V]\right| \\ &\leq \|P_h^o\|_1\|V - V'\|_\infty + \left|\mathbb{E}_{P_h^o}[V'] - \mathbb{E}_{\widehat{P}_h^k}[V']\right| + \left\|\widehat{P}_h^k\right\|_1\|V - V'\|_\infty \\ &\leq \sup_{V'\in\mathcal{N}_\mathcal{V}(\epsilon)} \left|\mathbb{E}_{P_h^o}[V'] - \mathbb{E}_{\widehat{P}_h^k}[V']\right| + 2\epsilon, \end{aligned} \tag{C.45}$$

where the second inequality follows from the Holder's inequality.

Then with probability at least $1-\delta$, we have

$$\begin{aligned}
\sup_{\boldsymbol{\lambda}\in[0,H]}\left|\mathbb{E}_{\widehat{P}_h^k}\left[V_{h+1}^{k,\rho}-\boldsymbol{\lambda}\right]-\mathbb{E}_{P_h^o}\left[V_{h+1}^{k,\rho}-\boldsymbol{\lambda}\right]\right| &\le \sup_{\eta\in[0,H]}\left|\mathbb{E}_{P_h^o}\left[V(\boldsymbol{\lambda})\right]-\mathbb{E}_{\widehat{P}_h^k}\left[V(\boldsymbol{\lambda})\right]\right| \\
&\le \sup_{V\in\mathcal{N}_{\mathcal{V}}(\epsilon)}\left|\mathbb{E}_{P_h^o}[V]-\mathbb{E}_{\widehat{P}_h^k}[V]\right|+2\epsilon \qquad \text{(C.46)} \\
&\le H\sqrt{\frac{2S\ln(2SAHK|\mathcal{N}_{\mathcal{V}}(\epsilon)|/\delta)}{n_h^{k-1}\vee 1}}+2\epsilon \qquad \text{(C.47)} \\
&\le H\sqrt{\frac{2S^2\ln(6SAH^2K/\epsilon\delta)}{n_h^{k-1}\vee 1}}+2\epsilon \\
&= H\sqrt{\frac{2S^2\ln(12SAH^2K^2/\delta)}{n_h^{k-1}\vee 1}}+\frac{1}{K}, \qquad \text{(C.48)}
\end{aligned}$$

for any $(k,h,s,a)\in[K]\times[H]\times\mathcal{S}\times\mathcal{A}$, where (C.46) follows from (C.45), (C.47) is from (C.22) and a union bound, we set $\epsilon=1/2K$ in (C.48).

We denote $V(\boldsymbol{\lambda})=V_{h+1}^{k,\rho}-\boldsymbol{\lambda}\in[-H,H]$ and $\mathcal{V}=\left\{V\in\mathbb{R}^S:\|V\|_\infty\le H\right\}$. To bound term (iii) in (C.43), we create a ϵ-net $\mathcal{N}_{\mathcal{V}}(\epsilon)$ for $\mathcal{V}$. From Lemma E.4, it holds that $\ln|\mathcal{N}_{\mathcal{V}}(\epsilon)|\le|S|\cdot\ln(3H/\epsilon)$.

Therefore, by the definition of $\mathcal{N}_{\mathcal{V}}(\epsilon)$, for any fixed V, there exists a $V'\in\mathcal{N}_{\mathcal{V}}(\epsilon)$ such that $\|V-V'\|_\infty\le\epsilon$, that is

$$\begin{aligned}
&\left|\mathrm{Var}_{P_h^o}(V)-\mathrm{Var}_{\widehat{P}_h^k}(V)\right| \\
\le&\left|\mathrm{Var}_{P_h^o}(V)-\mathrm{Var}_{P_h^o}(V')\right|+\left|\mathrm{Var}_{P_h^o}(V')-\mathrm{Var}_{\widehat{P}_h^k}(V')\right|+\left|\mathrm{Var}_{\widehat{P}_h^k}(V')-\mathrm{Var}_{\widehat{P}_h^k}(V)\right| \\
\le&\left|\mathbb{E}_{P_h^o}\left[V^2-V'^2\right]\right|+\left|\mathbb{E}_{P_h^o}^2[V]-\mathbb{E}_{P_h^o}^2[V']\right|+\left|\mathbb{E}_{\widehat{P}_h^k}\left[V'^2-V^2\right]\right|+\left|\mathbb{E}_{\widehat{P}_h^k}^2[V']-\mathbb{E}_{\widehat{P}_h^k}^2[V]\right| \\
&+\left|\mathrm{Var}_{P_h^o}(V')-\mathrm{Var}_{\widehat{P}_h^k}(V')\right| \\
\le&\left|\mathrm{Var}_{P_h^o}(V')-\mathrm{Var}_{\widehat{P}_h^k}(V')\right|+8H\epsilon \\
\le&\sup_{V'\in N_{\mathcal{V}}}\left|\mathrm{Var}_{P_h^o}(V')-\mathrm{Var}_{\widehat{P}_h^k}(V')\right|+8H\epsilon. \qquad \text{(C.49)}
\end{aligned}$$

For any fixed V, following the same analysis as (C.22), we have

$$\begin{aligned}
\left|\mathrm{Var}_{P_h^o}(V)-\mathrm{Var}_{\widehat{P}_h^k}(V)\right| &= \left|\left(\mathbb{E}_{P_h^o}[V^2]-\mathbb{E}_{P_h^o}^2[V]\right)-\left(\mathbb{E}_{\widehat{P}_h^k}[V^2]-\mathbb{E}_{\widehat{P}_h^k}^2[V]\right)\right| \\
&\le \left|\mathbb{E}_{P_h^o}[V^2]-\mathbb{E}_{\widehat{P}_h^k}[V^2]\right|+\left|\mathbb{E}_{P_h^o}^2[V]-\mathbb{E}_{\widehat{P}_h^k}^2[V]\right| \\
&\le H^2\sqrt{\frac{2S\ln(2/\delta)}{n_h^{k-1}\vee 1}}+\left(\mathbb{E}_{P_h^o}[V]+\mathbb{E}_{\widehat{P}_h^k}[V]\right)\cdot\left|\mathbb{E}_{P_h^o}[V]-\mathbb{E}_{\widehat{P}_h^k}[V]\right| \\
&\le H^2\sqrt{\frac{2S\ln(2/\delta)}{n_h^{k-1}\vee 1}}+2H^2\sqrt{\frac{2S\ln(2/\delta)}{n_h^{k-1}\vee 1}} \\
&\le 3H^2\sqrt{\frac{2S\ln(2/\delta)}{n_h^{k-1}\vee 1}} \qquad \text{(C.50)}
\end{aligned}$$

with probability at least $1-\delta$.

Then with probability at least $1-\delta$, we have

$$\begin{aligned}
\sup_{\boldsymbol{\lambda}\in[0,H]} \left|\mathrm{Var}_{\widehat{P}_h^k}\left(V_{h+1}^{k,\beta,\eta}-\boldsymbol{\lambda}\right)-\mathrm{Var}_{P_h^o}\left(V_{h+1}^{k,\beta,\eta}-\boldsymbol{\lambda}\right)\right| &\le \sup_{\boldsymbol{\lambda}\in[0,H]} \left|\mathrm{Var}_{P_h^o}(V(\boldsymbol{\lambda}))-\mathrm{Var}_{\widehat{P}_h^k}(V(\boldsymbol{\lambda}))\right| \\
&\le \sup_{V\in\mathcal{N}_{\mathcal{V}}(\epsilon)} \left|\mathrm{Var}_{P_h^o}(V)-\mathrm{Var}_{\widehat{P}_h^k}(V)\right|+8H\epsilon \quad \text{(C.51)} \\
&\le 3H^2\sqrt{\frac{2S\ln(2SAHK|\mathcal{N}_{\mathcal{V}}|/\delta)}{n_h^{k-1}\vee 1}}+8H\epsilon \quad \text{(C.52)} \\
&\le 3H^2\sqrt{\frac{2S^2\ln(6SAH^2K/\epsilon\delta)}{n_h^{k-1}\vee 1}}+8H\epsilon \\
&= 3H^2\sqrt{\frac{2S^2\ln(48SAH^3K^2/\delta)}{n_h^{k-1}\vee 1}}+\frac{1}{K}, \quad \text{(C.53)}
\end{aligned}$$

for any $(k,h,s,a)\in[K]\times[H]\times\mathcal{S}\times\mathcal{A}$, where (C.51) follows from (C.49), (C.52) is from (C.50) and a union bound, we set $\epsilon=\frac{1}{8HK}$ in (C.53).

Apply the union bound again and combine (C.43) with (C.44), (C.48), (C.53), the definition of bonus and induction assumption. With probability at least $1-3\delta$, we have $Q_h^{k,\rho}(s,a)\ge Q_h^{\pi,\rho}(s,a)$ for any $(k,h,s,a)\in[K]\times[H]\times\mathcal{S}\times\mathcal{A}$. This completes the proof. □

Lemma C.11. If we set the bonus term to be the same as in Lemma C.10, then for any $(k,s,a)\in[K]\times\mathcal{S}\times\mathcal{A}$, with probability at least $1-\delta$, the sum of estimation errors can be bounded as

$$Q_1^{k,\beta,\eta}(s,a)-Q_1^{\pi^k,\beta,\eta}(s,a)\le 2\cdot\mathbb{E}_{P^{w,k},\pi^k}\left[\mathrm{bonus}_h^k\right],$$

Proof. From the proof of Lemma C.10, we see that with probability at least $1-\delta$, for any $(k,s,a)\in[K]\times\mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}
&\left|\widehat{r}_h^k(s,a)-r_h(s,a)\right|+\Big|\inf_{P\in\Delta(S)}\left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right]+\beta\chi^2(P\|\widehat{P}_h^k)\right)(s,a) \\
&\quad-\inf_{P\in\Delta(S)}\left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right]+\beta\chi^2(P\|P_h^o)\right)(s,a)\Big|\le\mathrm{bonus}_h^k(s,a). \quad \text{(C.54)}
\end{aligned}$$

Recall that we define $P_h^{w,k}=\operatorname{argmin}_{P\in\Delta(S)}\left(\mathbb{E}_P\left[V_{h+1}^{\pi^k,\beta,\eta}\right]+\beta\chi^2(P\|P_h^o)\right)$ as the worst-case transition in Definition 5.1, we have

$$\begin{aligned}
&Q_h^{k,\beta,\eta}-Q_h^{\pi^k,\beta,\eta} \\
&\le \mathrm{bonus}_h^k+\widehat{r}_h^k+\inf_{P\in\Delta(S)}\left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right]+\beta\chi^2(P\|\widehat{P}_h^k)\right)-r_h-\inf_{P\in\Delta(S)}\left(\mathbb{E}_P\left[V_{h+1}^{\pi^k,\beta,\eta}\right]+\beta\chi^2(P\|P_h^o)\right) \\
&\le 2\mathrm{bonus}_h^k+\inf_{P\in\Delta(S)}\left(\mathbb{E}_P\left[V_{h+1}^{k,\beta,\eta}\right]+\beta\chi^2(P\|P_h^o)\right)-\inf_{P\in\Delta(S)}\left(\mathbb{E}_P\left[V_{h+1}^{\pi^k,\beta,\eta}\right]+\beta\chi^2(P\|P_h^o)\right) \quad \text{(C.55)} \\
&= 2\mathrm{bonus}_h^k+\mathbb{E}_{P_h^{w,k},\pi^k}\left[Q_{h+1}^{k,\beta,\eta}-Q_{h+1}^{\pi^k,\beta,\eta}\right]. \quad \text{(C.56)}
\end{aligned}$$

where (C.55) uses (C.54). Apply (C.56) recursively, we can obtain the result. □

Proof of Lemma C.1 in χ^2 Setting. The result directly follows from combining Lemma C.10 and Lemma C.11, along with applying a union bound. □

D Proofs of the Results in Linear Setting

D.1 Proof of the Dynamic Programming Principle

Proof of Proposition 6.2. We prove a stronger version of the proposition. Specifically, there exists transition weights $\{\widetilde{\boldsymbol{\mu}}_t\}_{t=1}^H$ together with $\widetilde{P}_t = \langle \boldsymbol{\phi}, \widetilde{\boldsymbol{\mu}}_t \rangle$, such that for all $(h, s) \in [H] \times \mathcal{S}$,

$$V_h^{\pi,\beta,\eta}(s) = \mathbb{E}_{\{\widetilde{P}_t\}_{t=h}^H}\left[\sum_{t=h}^H \left(r_t(s_t,a_t) - \Omega_{s_t}^\eta(\pi_t) + \beta\langle \boldsymbol{\phi}(s_t,a_t), \mathbf{D}(\widetilde{\boldsymbol{\mu}}_t \| \boldsymbol{\mu}_t^o)\rangle\right) \middle| s_h = s, \pi_{h:H}\right]. \tag{D.1}$$

We prove this statement by induction. For the base case $h = H$, the claim holds trivially since no transition kernels are involved. For the inductive step, assume the conclusion holds for step $h+1$, meaning that there exists $\{\widetilde{r}_t\}_{t=1}^H$ and $\{\widetilde{P}_t\}_{t=h+1}^H$ such that

$$V_{h+1}^{\pi,\beta,\eta}(s) = \mathbb{E}_{\{\widetilde{P}_t\}_{t=h+1}^H}\left[\sum_{t=h+1}^H \left(r_t(s_t,a_t) - \Omega_{s_t}^\eta(\pi_t) + \beta\langle \boldsymbol{\phi}(s_t,a_t), \mathbf{D}(\widetilde{\boldsymbol{\mu}}_t \| \boldsymbol{\mu}_t^o)\rangle\right) \middle| s_{h+1} = s, \pi_{h+1:H}\right]. \tag{D.2}$$

We now prove (D.1) for the case of h. For $Q_h^{\pi,\beta,\eta}(s,a)$, on the one hand, we upper bound it as

$$\begin{aligned}
&Q_h^{\pi,\beta,\eta}(s,a) \\
&= \inf_{\substack{\boldsymbol{\mu}_t \in \Delta(S)^d \\ P_t = \langle \boldsymbol{\phi}, \boldsymbol{\mu}_t\rangle}} \mathbb{E}_{\{P_t\}_{t=h}^H}\left[\sum_{t=h}^H \left(r_t(s_t,a_t) + \beta\langle \boldsymbol{\phi}(s_t,a_t), \mathbf{D}(\boldsymbol{\mu}_t \| \boldsymbol{\mu}_t^o)\rangle\right) - \sum_{t=h+1}^H \Omega_{s_t}^\eta(\pi_t) \middle| s_h = s, a_h = a, \pi_{h+1:H}\right] \\
&= r_h(s,a) + \inf_{\substack{\boldsymbol{\mu}_t \in \Delta(S)^d \\ P_t = \langle \boldsymbol{\phi}, \boldsymbol{\mu}_t\rangle}} \left\{\beta\langle \boldsymbol{\phi}(s_h,a_h), \mathbf{D}(\boldsymbol{\mu}_h \| \boldsymbol{\mu}_h^o)\rangle + \int_{\mathcal{S}} P_h(\mathrm{d}s'|s,a)\mathbb{E}_{\{P_t\}_{t=h+1}^H}\Bigg[\right. \\
&\qquad \left.\sum_{t=h+1}^H \left(r_t(s_t,a_t) - \Omega_{s_t}^\eta(\pi_t) + \beta\langle \boldsymbol{\phi}(s_t,a_t), \mathbf{D}(\boldsymbol{\mu}_t \| \boldsymbol{\mu}_t^o)\rangle\right) \middle| s_{h+1} = s', \pi_{h+1:H}\Bigg]\right\} \\
&\le r_h(s,a) + \inf_{\substack{\boldsymbol{\mu}_h \in \Delta(S)^d \\ P_h = \langle \boldsymbol{\phi}, \boldsymbol{\mu}_h\rangle}} \left\{\beta\langle \boldsymbol{\phi}(s_h,a_h), \mathbf{D}(\boldsymbol{\mu}_h \| \boldsymbol{\mu}_h^o)\rangle + \int_{\mathcal{S}} P_h(\mathrm{d}s'|s,a)\mathbb{E}_{\{\widetilde{P}_t\}_{t=h+1}^H}\Bigg[\right. \\
&\qquad \left.\sum_{t=h+1}^H \left(r_t(s_t,a_t) - \Omega_{s_t}^\eta(\pi_t) + \beta\langle \boldsymbol{\phi}(s_t,a_t), \mathbf{D}(\widetilde{\boldsymbol{\mu}}_t \| \boldsymbol{\mu}_t^o)\rangle\right) \middle| s_{h+1} = s', \pi_{h+1:H}\Bigg]\right\} \qquad \text{(D.3)} \\
&= r_h(s,a) + \inf_{\substack{\boldsymbol{\mu}_h \in \Delta(S)^d \\ P_h = \langle \boldsymbol{\phi}, \boldsymbol{\mu}_h\rangle}} \left\{\beta\langle \boldsymbol{\phi}(s_h,a_h), \mathbf{D}(\boldsymbol{\mu}_h \| \boldsymbol{\mu}_h^o)\rangle + \mathbb{E}_{s' \sim P_h^o(\cdot|s,a)}\left[V_{h+1}^{\pi,\beta,\eta}(s')\right]\right\}, \qquad \text{(D.4)}
\end{aligned}$$

where (D.3) is because the definition of the infimum operator and (D.4) is from (D.2).

On the other hand, we can lower bound $Q_h^{\pi,\beta,\eta}(s,a)$ as

$$\begin{aligned}
&Q_h^{\pi,\beta,\eta}(s,a) \\
&= \inf_{\substack{\boldsymbol{\mu}_t \in \Delta(S)^d \\ P_t = \langle \boldsymbol{\phi}, \boldsymbol{\mu}_t\rangle}} \mathbb{E}_{\{P_t\}_{t=h}^H}\left[\sum_{t=h}^H \left(r_t(s_t,a_t) + \beta\langle \boldsymbol{\phi}(s_t,a_t), \mathbf{D}(\boldsymbol{\mu}_t \| \boldsymbol{\mu}_t^o)\rangle\right) - \sum_{t=h+1}^H \Omega_{s_t}^\eta(\pi_t) \middle| s_h = s, a_h = a, \pi_{h+1:H}\right] \\
&= r_h(s,a) + \inf_{\substack{\boldsymbol{\mu}_t \in \Delta(S)^d \\ P_t = \langle \boldsymbol{\phi}, \boldsymbol{\mu}_t\rangle}} \left\{\beta\langle \boldsymbol{\phi}(s_h,a_h), \mathbf{D}(\boldsymbol{\mu}_h \| \boldsymbol{\mu}_h^o)\rangle + \int_{\mathcal{S}} P_h(\mathrm{d}s'|s,a)\mathbb{E}_{\{P_t\}_{t=h+1}^H}\Bigg[\right.
\end{aligned}$$

$$\sum_{t=h+1}^{H}\Big(r_t(s_t,a_t)-\Omega^{\eta}_{s_t}(\pi_t)+\beta\big\langle\boldsymbol{\phi}(s_t,a_t),\mathbf{D}(\boldsymbol{\mu}_t\|\boldsymbol{\mu}^o_t)\big\rangle\Big)\Big|s_{h+1}=s',\pi_{h+1:H}\Big]\Big\}$$

$$\geq r_h(s,a)+\inf_{\substack{\boldsymbol{\mu}_h\in\Delta(S)^d\\P_h=\langle\boldsymbol{\phi},\boldsymbol{\mu}_h\rangle}}\Big\{\beta\big\langle\boldsymbol{\phi}(s_h,a_h),\mathbf{D}(\boldsymbol{\mu}_h\|\boldsymbol{\mu}^o_h)\big\rangle+\int_{\mathcal{S}}P_h(\mathrm{d}s'|s,a)\inf_{\substack{\boldsymbol{\mu}_t\in\Delta(S)^d\\P_t=\langle\boldsymbol{\phi},\boldsymbol{\mu}_t\rangle}}\mathbb{E}_{\{P_t\}_{t=h+1}^{H}}\Big[$$

$$\sum_{t=h+1}^{H}\Big(r_t(s_t,a_t)-\Omega^{\eta}_{s_t}(\pi_t)+\beta\big\langle\boldsymbol{\phi}(s_t,a_t),\mathbf{D}(\boldsymbol{\mu}_t\|\boldsymbol{\mu}^o_t)\big\rangle\Big)\Big|s_{h+1}=s',\pi_{h+1:H}\Big]\Big\} \tag{D.5}$$

$$=r_h(s,a)+\inf_{\substack{\boldsymbol{\mu}_h\in\Delta(S)^d\\P_h=\langle\boldsymbol{\phi},\boldsymbol{\mu}_h\rangle}}\Big\{\beta\big\langle\boldsymbol{\phi}(s_h,a_h),\mathbf{D}(\boldsymbol{\mu}_h\|\boldsymbol{\mu}^o_h)\big\rangle+\mathbb{E}_{s'\sim P^o_h(\cdot|s,a)}\big[V^{\pi,\beta,\eta}_{h+1}(s')\big]\Big\}, \tag{D.6}$$

where (D.5) is because $\inf\int f\geq\int\inf f$ and (D.6) is from the definition of $V^{\pi,\beta,\eta}_{h+1}(s')$.

By combining (D.4) and (D.6), we have

$$Q^{\pi,\beta,\eta}_h(s,a)=r_h(s,a)+\inf_{\substack{\boldsymbol{\mu}_h\in\Delta(S)^d\\P_h=\langle\boldsymbol{\phi},\boldsymbol{\mu}_h\rangle}}\Big\{\mathbb{E}_{s'\sim P^o_h(\cdot|s,a)}\big[V^{\pi,\beta,\eta}_{h+1}(s')\big]+\beta\big\langle\boldsymbol{\phi}(s,a),\mathbf{D}(\boldsymbol{\mu}_h\|\boldsymbol{\mu}^o_h)\big\rangle\Big\},$$

which implies (6.1).

Since $\Delta(\mathcal{S})$ is compact, there exists $\widetilde{\boldsymbol{\mu}}_h$ together with $\widetilde{P}_h=\langle\boldsymbol{\phi},\widetilde{\boldsymbol{\mu}}_h\rangle$, such that

$$Q^{\pi,\beta,\eta}_h(s,a)=r_h(s,a)+\mathbb{E}_{s'\sim\widetilde{P}_h(\cdot|s,a)}\big[V^{\pi,\beta,\eta}_{h+1}(s')\big]+\beta\big\langle\boldsymbol{\phi}(s,a),\mathbf{D}(\widetilde{\boldsymbol{\mu}}_h\|\boldsymbol{\mu}^o_h)\big\rangle. \tag{D.7}$$

This finishes the proof of (D.1).

For $V^{\pi,\beta,\eta}_h(s)$, on the one hand, we upper bound it as

$$V^{\pi,\beta,\eta}_h(s)$$

$$=\inf_{\substack{\boldsymbol{\mu}_t\in\Delta(S)^d\\P_t=\langle\boldsymbol{\phi},\boldsymbol{\mu}_t\rangle}}\mathbb{E}_{\{P_t\}_{t=h}^{H}}\Big[\sum_{t=h}^{H}\Big(r_t(s_t,a_t)-\Omega^{\eta}_{s_t}(\pi_t)+\beta\big\langle\boldsymbol{\phi}(s_t,a_t),\mathbf{D}(\boldsymbol{\mu}_t\|\boldsymbol{\mu}^o_t)\big\rangle\Big)\Big|s_h=s,\pi_{h:H}\Big]$$

$$=\inf_{\substack{\boldsymbol{\mu}_t\in\Delta(S)^d\\P_t=\langle\boldsymbol{\phi},\boldsymbol{\mu}_t\rangle}}\Big\{-\Omega^{\eta}_s(\pi_h)+\sum_{a\in\mathcal{A}}\pi_h(a|s)\mathbb{E}_{\{P_t\}_{t=h}^{H}}\Big[$$

$$\sum_{t=h}^{H}\Big(r_t(s_t,a_t)+\beta\big\langle\boldsymbol{\phi}(s_t,a_t),\mathbf{D}(\boldsymbol{\mu}_t\|\boldsymbol{\mu}^o_t)\big\rangle\Big)-\sum_{t=h+1}^{H}\Omega^{\eta}_{s_t}(\pi_t)\Big|s_h=s,a_h=a,\pi_{h+1:H}\Big]\Big\}$$

$$\leq-\Omega^{\eta}_s(\pi_h)+\sum_{a\in\mathcal{A}}\pi_h(a|s)\mathbb{E}_{\{\widetilde{P}_t\}_{t=h}^{H}}\Big[$$

$$\sum_{t=h}^{H}\Big(r_t(s_t,a_t)+\beta\big\langle\boldsymbol{\phi}(s_t,a_t),\mathbf{D}(\widetilde{\boldsymbol{\mu}}_t\|\boldsymbol{\mu}^o_t)\big\rangle\Big)-\sum_{t=h+1}^{H}\Omega^{\eta}_{s_t}(\pi_t)\Big|s_h=s,a_h=a,\pi_{h+1:H}\Big] \tag{D.8}$$

$$=-\Omega^{\eta}_s(\pi_h)+\sum_{a\in\mathcal{A}}\pi_h(a|s)Q^{\pi,\beta,\eta}_h(s,a), \tag{D.9}$$

where (D.8) is because the definition of the infimum operator and (D.9) is from (D.7) and (D.2).

On the other hand, we can lower bound $V^{\pi,\beta,\eta}_h(s)$ as

$$V^{\pi,\beta,\eta}_h(s)$$

$$
\begin{aligned}
&= \inf_{\substack{\boldsymbol{\mu}_t \in \Delta(S)^d \\ P_t = \langle \boldsymbol{\phi}, \boldsymbol{\mu}_t \rangle}} \mathbb{E}_{\{P_t\}_{t=h}^H} \left[\sum_{t=h}^H \Big(r_t(s_t, a_t) - \Omega_{s_t}^\eta(\pi_t) + \beta \big\langle \boldsymbol{\phi}(s_t, a_t), \mathbf{D}(\boldsymbol{\mu}_t \| \boldsymbol{\mu}_t^o) \big\rangle \Big) \middle| s_h = s, \pi_{h:H} \right] \\
&= \inf_{\substack{\boldsymbol{\mu}_t \in \Delta(S)^d \\ P_t = \langle \boldsymbol{\phi}, \boldsymbol{\mu}_t \rangle}} \Bigg\{ - \Omega_s^\eta(\pi_h) + \sum_{a \in \mathcal{A}} \pi_h(a|s) \mathbb{E}_{\{P_t\}_{t=h}^H} \Bigg[\\
&\qquad \sum_{t=h}^H \Big(r_t(s_t, a_t) + \beta \big\langle \boldsymbol{\phi}(s_t, a_t), \mathbf{D}(\boldsymbol{\mu}_t \| \boldsymbol{\mu}_t^o) \big\rangle \Big) - \sum_{t=h+1}^H \Omega_{s_t}^\eta(\pi_t) \Big| s_h = s, a_h = a, \pi_{h+1:H} \Bigg] \Bigg\} \\
&\geq - \Omega_s^\eta(\pi_h) + \sum_{a \in \mathcal{A}} \pi_h(a|s) \inf_{\substack{\boldsymbol{\mu}_t \in \Delta(S)^d \\ P_t = \langle \boldsymbol{\phi}, \boldsymbol{\mu}_t \rangle}} \mathbb{E}_{\{P_t\}_{t=h}^H} \Bigg[\\
&\qquad \sum_{t=h}^H \Big(r_t(s_t, a_t) + \beta \big\langle \boldsymbol{\phi}(s_t, a_t), \mathbf{D}(\boldsymbol{\mu}_t \| \boldsymbol{\mu}_t^o) \big\rangle \Big) - \sum_{t=h+1}^H \Omega_{s_t}^\eta(\pi_t) \Big| s_h = s, a_h = a, \pi_{h+1:H} \Bigg] \qquad \text{(D.10)} \\
&= - \Omega_s^\eta(\pi_h) + \sum_{a \in \mathcal{A}} \pi_h(a|s) Q_h^{\pi,\beta,\eta}(s, a), \qquad \text{(D.11)}
\end{aligned}
$$

where (D.10) is because $\inf \sum f \geq \sum \inf f$ and (D.11) is from the definition of $Q_h^{\pi,\beta,\eta}(s,a)$.

By combining (D.9) and (D.11), we have

$$
V_h^{\pi,\beta,\eta}(s) = -\Omega_s^\eta(\pi_h) + \sum_{a \in \mathcal{A}} \pi_h(a|s) Q_h^{\pi,\beta,\eta}(s, a),
$$

which implies (6.2). □

D.2 Proofs of the Main Results

To establish Theorems 6.6 and 6.7, we first introduce Lemma D.1, which bounds the regret by the sum of bonus terms under the expectation defined by the worst-case transition.

Lemma D.1. Under Assumption 6.5 and Assumption 6.1, for each f-divergence setting and any $\delta \in (0, 1/2)$, if we choose $\lambda = 1$ and set c according to Theorem 6.6 in Algorithm 2, then with probability at least $1 - 2\delta$, the regret of Algorithm 2 satisfies

$$
\mathrm{Regret}(K) \leq 2c \sum_{k=1}^K \sum_{h=1}^H \sum_{i=1}^d \mathbb{E}_{P^{w,k}, \pi^k} \Big[\big\| \phi_i(s, a) \mathbb{1}_i \big\|_{(\Lambda_h^k)^{-1}} \Big]. \qquad \text{(D.12)}
$$

Proof of Theorem 6.6. Recall that

$$
\Gamma_h^k(s, a) = c \sum_{i=1}^d \| \phi_i(s, a) \mathbb{1}_i \|_{(\Lambda_h^k)^{-1}}
$$

is the bonus assigned to the state-action pair (s, a) at episode k and step h. Due to the truncation in (6.3), we may assume without loss of generality that $\Gamma_h^k(s, a) \leq H$ for all $(k, h, s, a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$. Since otherwise we can simply truncate $\Gamma_h^k(s, a)$ at H, which does not affect the results.

Therefore, by Azuma-Hoeffding inequality, with probability at least $1 - \delta$, we have

$$
\left| \sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{P^o, \pi^k} \big[\Gamma_h^k(s, a) \big] - \sum_{k=1}^K \sum_{h=1}^H \Gamma_h^k(s_h^k, a_h^k) \right| \leq \sqrt{2KH^3 \ln(2/\delta)}.
$$

Now we further analyze (D.12) as

$$\begin{aligned}
\text{Regret}(K) &\le 2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{P^{w,k},\pi^k}\left[\Gamma_h^k(s,a)\right] \\
&\le 2C_{vr}\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{P^o,\pi^k}\left[\Gamma_h^k(s,a)\right] \qquad \text{(D.13)} \\
&\le 2C_{vr}\left(\sqrt{2KH^3\ln(2/\delta)}+\sum_{k=1}^{K}\sum_{h=1}^{H}\Gamma_h^k(s_h^k,a_h^k)\right) \\
&= 2C_{vr}\left(\sqrt{2KH^3\ln(2/\delta)}+c\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{d}\left\|\phi_i(s_h^k,a_h^k)\mathbb{1}_i\right\|_{(\Lambda_h^k)^{-1}}\right),
\end{aligned}$$

where (D.13) follows from Assumption 6.5. This finishes the proof. □

Proof of Theorem 6.7. Now we further analyze (D.12) as

$$\begin{aligned}
\text{Regret}(K) &\le 2c\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{d}\mathbb{E}_{P^{w,k},\pi^k}\left[\left\|\phi_i(s,a)\mathbb{1}_i\right\|_{(\Lambda_h^k)^{-1}}\right] \\
&\le 2c\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{P^{w,k},\pi^k}\left[\sum_{i=1}^{d}\phi_i(s,a)\sqrt{\lambda_{\max}\left((\Lambda_h^k)^{-1}\right)}\right] \qquad \text{(D.14)} \\
&= 2c\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{P^{w,k},\pi^k}\left[\sqrt{\lambda_{\max}\left((\Lambda_h^k)^{-1}\right)}\right] \\
&\le 2c\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\mathbb{E}_{P^{w,k},\pi^k}\left[\lambda_{\max}\left((\Lambda_h^k)^{-1}\right)\right]} \qquad \text{(D.15)} \\
&\le 2c\sqrt{K}\sum_{h=1}^{H}\sqrt{\sum_{k=1}^{K}\mathbb{E}_{P^{w,k},\pi^k}\left[\lambda_{\max}\left((\Lambda_h^k)^{-1}\right)\right]}, \qquad \text{(D.16)}
\end{aligned}$$

where (D.14) is because $\|x\|_A \le \sqrt{\lambda_{\max}(A)}\|x\|_2$, (D.15) is from Jensen's inequality, (D.16) is from Cauchy-Schwarz inequality.

Using Lemma E.8, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}_{P^{w,k},\pi^k}\left[\lambda_{\max}\left((\Lambda_h^k)^{-1}\right)\right] \\
&= \sum_{k=1}^{K}\mathbb{E}_{P^{w,k},\pi^k}\left[\frac{1}{\lambda_{\min}(\Lambda_h^k)}\right] \\
&\le \sum_{k=1}^{K}\mathbb{E}_{P^{w,k},\pi^k}\left[\frac{1}{\max\left\{\alpha(k-1)+\lambda-\sqrt{32k\log(dKH/\delta)},\lambda\right\}}\right] \\
&= \sum_{k=1}^{K}\frac{1}{\max\left\{\alpha(k-1)+\lambda-\sqrt{32k\log(dKH/\delta)},\lambda\right\}}
\end{aligned}$$

$$\leq \frac{128}{\alpha^2}\log\left(\frac{dHK}{\delta}\right) + \frac{2}{\alpha}\log(K). \tag{D.17}$$

Combining (D.16) and (D.17), we finish the proof. □

D.3 Proofs of the Technical Lemmas

Before proving Lemma D.1 in different settings, we state a technical lemma about the covering number.

Lemma D.2. For any $h \in [H]$, let $\mathcal{V}_h$ denote a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with the following form

$$V_h(s; w, \beta, \Lambda) = \left[\max_{\pi\in\Pi}\left\{\left\langle \pi(s,a), \boldsymbol{\phi}(s,a)^\top \boldsymbol{w} + \beta\sum_{i=1}^d \|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda^{-1}}\right\rangle_{\mathcal{A}} - \sigma_{\mathcal{R}_h}\big(-\pi_h(\cdot|s)\big)\right\}\right]_{[0,H-h+1]},$$

where the parameters (w, β, Λ) satisfy $\|w\| \leq L$, $\beta \in [0, B]$, $\lambda_{\min}(\Lambda) \geq \lambda$. Let $\mathcal{N}_h(\epsilon; \mathcal{V}_h)$ be the ϵ-covering number of $\mathcal{V}_h$ with respect to the distance $\operatorname{dist}(V_1, V_2) = \sup_{s\in\mathcal{S}} |V_1(s) - V_2(s)|$. Then

$$\log \mathcal{N}_h(\epsilon) \leq d\log(1 + 4L|\mathcal{A}|/\epsilon) + d^2\log\big(1 + 8d^{1/2}B^2|\mathcal{A}|^2/(\lambda\epsilon^2)\big).$$

Proof. The argument is similar to that of Liu and Xu (2024a, Lemma D.3). Denoting $A = \beta^2\Lambda^{-1}$, we have

$$V_h(s; w, \beta, \Lambda) = \left[\max_{\pi\in\Pi}\left\{\left\langle \pi(s,a), \boldsymbol{\phi}(s,a)^\top \boldsymbol{w} + \beta\sum_{i=1}^d \|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda^{-1}}\right\rangle_{\mathcal{A}} - \sigma_{\mathcal{R}_h}\big(-\pi_h(\cdot|s)\big)\right\}\right]_{[0,H-h+1]},$$

for $\|w\| \leq L$ and $\|A\| \leq B^2\lambda^{-1}$. For any two functions $V_1, V_2 \in \mathcal{V}_h$, let them take the form above with parameters $(\boldsymbol{w}_1, A_1)$ and $(\boldsymbol{w}_2, A_2)$ respectively. Since $\min\{\,\cdot\,, H-h+1\}$ is a concentration map, we have

$$\begin{aligned}
&\operatorname{dist}(V_1, V_2)\\
&\leq \sup_{s,\pi}\left|\left\{\left\langle \pi(s,a), \boldsymbol{\phi}(s,a)^\top \boldsymbol{w}_1 + \sum_{i=1}^d \|\phi_i(s,a)\mathbb{1}_i\|_{A_1}\right\rangle_{\mathcal{A}} - \sigma_{\mathcal{R}_h}\big(-\pi_h(\cdot|s)\big)\right\}\right.\\
&\qquad\left. - \left\{\left\langle \pi(s,a), \boldsymbol{\phi}(s,a)^\top \boldsymbol{w}_2 + \sum_{i=1}^d \|\phi_i(s,a)\mathbb{1}_i\|_{A_2}\right\rangle_{\mathcal{A}} - \sigma_{\mathcal{R}_h}\big(-\pi_h(\cdot|s)\big)\right\}\right|\\
&\leq \sup_{s,\pi}\left|\left\langle \pi(s,a), \left(\boldsymbol{\phi}(s,a)^\top \boldsymbol{w}_1 + \sum_{i=1}^d \|\phi_i(s,a)\mathbb{1}_i\|_{A_1}\right) - \left(\boldsymbol{\phi}(s,a)^\top \boldsymbol{w}_2 + \sum_{i=1}^d \|\phi_i(s,a)\mathbb{1}_i\|_{A_2}\right)\right\rangle_{\mathcal{A}}\right|\\
&\leq \sup_{s,\pi}\left|\big\langle \pi(s,a), \boldsymbol{\phi}(s,a)^\top(\boldsymbol{w}_1 - \boldsymbol{w}_2)\big\rangle\right| + \sup_{s,\pi}\left|\left\langle \pi(s,a), \sum_{i=1}^d \|\phi_i(s,a)\mathbb{1}_i\|_{A_1} - \sum_{i=1}^d \|\phi_i(s,a)\mathbb{1}_i\|_{A_2}\right\rangle_{\mathcal{A}}\right|\\
&\leq \sup_{s,\pi}\left|\big\langle \pi(s,a), \boldsymbol{\phi}(s,a)^\top(\boldsymbol{w}_1 - \boldsymbol{w}_2)\big\rangle\right| + \sup_{s,\pi}\left|\left\langle \pi(s,a), \sum_{i=1}^d \|\phi_i(s,a)\mathbb{1}_i\|_{A_1 - A_2}\right\rangle_{\mathcal{A}}\right|
\end{aligned} \tag{D.18}$$

$$\leq |\mathcal{A}|\Big(\|\boldsymbol{w}_1 - \boldsymbol{w}_2\| + \sqrt{\|A_1 - A_2\|_F}\Big),$$

where (D.18) is because $|\sqrt{x} - \sqrt{y}| \leq \sqrt{|x-y|}$ for $x, y \geq 0$. For matrices, $\|\cdot\|$ and $\|\cdot\|_F$ denote the matrix operator norm and Frobenius norm respectively.

Let $\mathcal{C}_{\boldsymbol{w}}$ be an $\epsilon/(2|\mathcal{A}|)$-cover of $\big\{\boldsymbol{w} \in \mathbb{R}^d \,\big|\, \|\boldsymbol{w}\|_2 \leq L\big\}$ with respect to the 2-norm and $\mathcal{C}_A$ be an $\epsilon^2/(4|\mathcal{A}|^2)$-cover of $\big\{A \in \mathbb{R}^{d\times d} \,\big|\, \|A\|_F \leq d^{1/2}B^2\lambda^{-1}\big\}$ with respect to the Frobenius norm. By Lemma E.6, we have

$$\log \mathcal{N}_\epsilon \leq \log|\mathcal{C}_{\boldsymbol{w}}| + \log|\mathcal{C}_A| \leq d\log(1 + 4L|\mathcal{A}|/\epsilon) + d^2\log\big(1 + 8d^{1/2}B^2|\mathcal{A}|^2/(\lambda\epsilon^2)\big).$$

This finishes the proof. □

Next, we present the proof of Lemma D.1 is various f-divergence settings. For convenience, we also write $P^{w,k} := P^{w,\pi^k}$, $\mathrm{d}^k := \mathrm{d}^{\pi^k}$ and $\mathrm{q}^k := \mathrm{q}^{\pi^k}$.

D.3.1 Proof of Lemma D.1 in TV Setting

Lemma D.3 (Dual formulation)**.** (Tang et al., 2025, Proposition 4.2) For the optimization problem $\inf_{\boldsymbol{\mu}\in\Delta(\mathcal{S})}\big(\mathbb{E}_{s\sim\boldsymbol{\mu}}[V_{h+1}] + \beta\mathrm{TV}(\boldsymbol{\mu}\|\boldsymbol{\mu}^o)\big)$, we have its dual formulation as follows

$$\inf_{\boldsymbol{\mu}\in\Delta(\mathcal{S})}\big(\mathbb{E}_{s\sim\boldsymbol{\mu}}[V_{h+1}] + \beta\mathrm{TV}(\boldsymbol{\mu}\|\boldsymbol{\mu}^o)\big) = \mathbb{E}_{s\sim\boldsymbol{\mu}^o}\big[V_{h+1}(s)\big]_{\min_{s'\in\mathcal{S}} V_{h+1}(s')+\beta}.$$

Lemma D.4 (Optimism)**.** If we set the bonus term as follows

$$\begin{aligned}
\Gamma_h^k(s,a) &= c_{\mathrm{TV}}\sum_{i=1}^d \|\phi_i(s,a)\mathbb{1}_i\|_{(\Lambda_h^k)^{-1}},\\
\text{where}\quad & c_{\mathrm{TV}} = Hd\cdot\xi_{\mathrm{TV}},\\
\text{where}\quad & \xi_{\mathrm{TV}} = 720 + 3\sqrt{40\log\big(96K^{13/2}H|\mathcal{A}|^3/\delta\big)},
\end{aligned}$$

then for any policy π and any $(k,h,s,a) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}$, with probability at least $1-\delta$, we have $Q_h^{k,\beta,\eta}(s,a) \geq Q_h^{\pi,\beta,\eta}(s,a)$. Specially, by setting $\pi = \pi^*$, we have $Q_h^{k,\beta,\eta}(s,a) \geq Q_h^{\pi^*,\beta,\eta}(s,a)$.

Proof. We prove this by induction. First, $h = H+1$ holds trivially since $Q_{H+1}^{k,\beta,\eta}(s,a) = 0 = Q_{H+1}^{\pi,\beta,\eta}(s,a)$.

Assume $Q_{h+1}^{k,\beta,\eta}(s,a) \geq Q_{h+1}^{\pi,\beta,\eta}(s,a)$ holds, due to the choice of π_{h+1}^k in (4.3), we have

$$\begin{aligned}
V_{h+1}^{k,\beta,\eta}(s) &= \big\langle Q_{h+1}^{k,\beta,\eta}(s,\cdot), \pi_{h+1}^k(\cdot|s)\big\rangle - \Omega_s^\eta(\pi_{h+1}^k)\\
&\geq \big\langle Q_{h+1}^{k,\beta,\eta}(s,\cdot), \pi_{h+1}(\cdot|s)\big\rangle - \Omega_s^\eta(\pi_{h+1})\\
&\geq \big\langle Q_{h+1}^{\pi,\beta,\eta}(s,\cdot), \pi_{h+1}(\cdot|s)\big\rangle - \Omega_s^\eta(\pi_{h+1}) = V_{h+1}^{\pi,\beta,\eta}(s).
\end{aligned}$$

Recall that we denote $Q_h^{k,\beta,\eta}$ as the optimistic estimation in k-th episode, that is,

$$Q_h^{k,\beta,\eta}(s,a) \leftarrow \min\big\{\boldsymbol{\phi}(s,a)^\top\big(\boldsymbol{\theta}_h^k + \boldsymbol{w}_h^k\big) + \Gamma_h^k(s,a), H-h+1\big\}. \tag{D.19}$$

If $Q_h^{k,\beta,\eta}(s,a) = H-h+1$, then it follows immediately that

$$Q_h^{k,\beta,\eta}(s,a) = H-h+1 \geq Q_h^{\pi,\beta,\eta}(s,a)$$

by the definition of $Q_h^{\pi,\beta,\eta}(s,a)$. Otherwise, we can infer that

$$\begin{aligned}
&Q_h^{k,\beta,\eta} - Q_h^{\pi,\beta,\eta} \\
&= \Gamma_h^k + \boldsymbol{\phi}^\top\left(\boldsymbol{\theta}_h^k + \boldsymbol{w}_h^k\right) - \boldsymbol{\phi}^\top\left[\boldsymbol{\theta}_h + \inf_{\mu\in\Delta(\mathcal{S})}\left(\mathbb{E}_{s\sim\mu}[V_{h+1}^{\pi,\beta,\eta}] + \beta\mathrm{TV}(\mu\|\mu^o)\right)\right] \\
&\geq \Gamma_h^k + \boldsymbol{\phi}^\top\left(\boldsymbol{\theta}_h^k + \boldsymbol{w}_h^k\right) - \boldsymbol{\phi}^\top\left[\boldsymbol{\theta}_h + \inf_{\mu\in\Delta(\mathcal{S})}\left(\mathbb{E}_{s\sim\mu}[V_{h+1}^{k,\beta,\eta}] + \beta\mathrm{TV}(\mu\|\mu^o)\right)\right] \qquad \text{(D.20)} \\
&= \Gamma_h^k + \left\langle \boldsymbol{\phi}, \boldsymbol{\theta}_h^k + \widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}\left[V_{h+1}^{k,\beta,\eta}\right]_{\min_{s'} V_{h+1}^{k,\beta,\eta}(s')+\beta} - \boldsymbol{\theta}_h - \mathbb{E}_{s\sim\boldsymbol{\mu}^o}\left[V_{h+1}^{k,\beta,\eta}\right]_{\min_{s'} V_{h+1}^{k,\beta,\eta}(s')+\beta}\right\rangle \qquad \text{(D.21)} \\
&\geq \Gamma_h^k - \left\langle \boldsymbol{\phi}, \left|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h\right| + \left|\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}\left[V_{h+1}^{k,\beta,\eta}\right]_{\min_{s'} V_{h+1}^{k,\beta,\eta}(s')+\beta} - \mathbb{E}_{s\sim\boldsymbol{\mu}^o}\left[V_{h+1}^{k,\beta,\eta}\right]_{\min_{s'} V_{h+1}^{k,\beta,\eta}(s')+\beta}\right|\right\rangle \\
&= \Gamma_h^k - \sum_{i=1}^d \underbrace{\phi_i \mathbb{1}_i^\top\left|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h\right|}_{\text{(i)}} - \sum_{i=1}^d \underbrace{\phi_i \mathbb{1}_i^\top\left|\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}\left[V_{h+1}^{k,\beta,\eta}\right]_{\min_{s'} V_{h+1}^{k,\beta,\eta}(s')+\beta} - \mathbb{E}_{s\sim\boldsymbol{\mu}^o}\left[V_{h+1}^{k,\beta,\eta}\right]_{\min_{s'} V_{h+1}^{k,\beta,\eta}(s')+\beta}\right|}_{\text{(ii)}}, \qquad \text{(D.22)}
\end{aligned}$$

where (D.20) follows from the induction assumption, and (D.21) is obtained from the algorithm update formulation and the dual formulation.

We introduce the following notation for the sake of convenience

$$\mathrm{err}_h^\tau(f) = \mathbb{E}_{s'\sim\mathbb{P}_h^o(\cdot|s_h^\tau,a_h^\tau)}[f(s')] - f(s_{h+1}^\tau). \qquad \text{(D.23)}$$

For term (i) in (D.22), we have

$$\begin{aligned}
&\left|\phi_i \mathbb{1}_i^\top\left(\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h\right)\right| \\
&= \left|\phi_i \mathbb{1}_i^\top\left(\Lambda_h^k\right)^{-1}\left(\sum_{\tau=1}^{k-1}\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\cdot r_h^\tau\right)\right. \\
&\qquad \left. - \phi_i \mathbb{1}_i^\top\left(\Lambda_h^k\right)^{-1}\left(\sum_{\tau=1}^{k-1}\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\boldsymbol{\phi}(s_h^\tau,a_h^\tau)^\top + \lambda\cdot I\right)\cdot\boldsymbol{\theta}_h\right| \\
&= \lambda\left|\phi_i \mathbb{1}_i^\top\left(\Lambda_h^k\right)^{-1}\cdot\boldsymbol{\theta}_h\right| \qquad \text{(D.24)} \\
&\leq \lambda\|\phi_i\mathbb{1}_i\|_{(\Lambda_h^k)^{-1}}\cdot\|\boldsymbol{\theta}_h\|_{(\Lambda_h^k)^{-1}} \qquad \text{(D.25)} \\
&\leq \lambda\|\phi_i\mathbb{1}_i\|_{(\Lambda_h^k)^{-1}}\cdot\lambda_{\max}\left(\left(\Lambda_h^k\right)^{-1}\right)^{\frac{1}{2}}\|\boldsymbol{\theta}_h\|_2 \qquad \text{(D.26)} \\
&\leq \lambda^{\frac{1}{2}}\sqrt{d}\cdot\|\phi_i\mathbb{1}_i\|_{(\Lambda_h^k)^{-1}},
\end{aligned}$$

where (D.24) is obtained from the algorithm update formulation and the definition of Λ_h^k, (D.25) follows from the Cauchy-Schwartz inequality, (D.26) is because $\|x\|_A \leq \sqrt{\lambda_{\max}(A)}\|x\|_2$.

For term (ii) in (D.22), we decompose it as follows.

$$\left|\phi_i \mathbb{1}_i^\top\left(\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}\left[V_{h+1}^{k,\beta,\eta}\right]_{\min_{s'} V_{h+1}^{k,\beta,\eta}(s')+\beta} - \mathbb{E}_{s\sim\boldsymbol{\mu}^o}\left[V_{h+1}^{k,\beta,\eta}\right]_{\min_{s'} V_{h+1}^{k,\beta,\eta}(s')+\beta}\right)\right|$$

$$
\begin{aligned}
&= \Bigg| \phi_i \mathbb{1}_i^\top \big(\Lambda_h^k\big)^{-1} \Bigg(\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \big[V_{h+1}^{k,\beta,\eta}(s_{h+1}^\tau)\big]_{\min_{s'} V_{h+1}^{k,\beta,\eta}(s')+\beta} \Bigg) \\
&\qquad - \phi_i \mathbb{1}_i^\top \big(\Lambda_h^k\big)^{-1} \Bigg(\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top + \lambda \cdot I \Bigg) \Big(\mathbb{E}_{s\sim\boldsymbol{\mu}^o} \big[V_{h+1}^{k,\beta,\eta}(s)\big]_{\min_{s'} V_{h+1}^{k,\beta,\eta}(s')+\beta} \Big) \Bigg| \qquad \text{(D.27)} \\
&= \Bigg| \phi_i \mathbb{1}_i^\top \big(\Lambda_h^k\big)^{-1} \Bigg(\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau \Big(\big[V_{h+1}^{k,\beta,\eta}(s)\big]_{\min_{s'} V_{h+1}^{k,\beta,\eta}(s')+\beta} \Big) \Bigg) \\
&\qquad + \lambda \phi_i \mathbb{1}_i^\top \big(\Lambda_h^k\big)^{-1} \Big(\mathbb{E}_{s\sim\boldsymbol{\mu}^o} \big[V_{h+1}^{k,\beta,\eta}(s)\big]_{\min_{s'} V_{h+1}^{k,\beta,\eta}(s')+\beta} \Big) \Bigg| \qquad \text{(D.28)} \\
&\le \Bigg| \phi_i \mathbb{1}_i^\top \big(\Lambda_h^k\big)^{-1} \Bigg(\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau \Big(\big[V_{h+1}^{k,\beta,\eta}(s)\big]_{\min_{s'} V_{h+1}^{k,\beta,\eta}(s')+\beta} \Big) \Bigg) \Bigg| \\
&\qquad + \lambda \Big| \phi_i \mathbb{1}_i^\top \big(\Lambda_h^k\big)^{-1} \Big(\mathbb{E}_{s\sim\boldsymbol{\mu}^o} \big[V_{h+1}^{k,\beta,\eta}(s)\big]_{\min_{s'} V_{h+1}^{k,\beta,\eta}(s')+\beta} \Big) \Big| \\
&\le \|\phi_i \mathbb{1}_i\|_{(\Lambda_h^k)^{-1}} \cdot \underbrace{\Bigg\| \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau \Big(\big[V_{h+1}^{k,\beta,\eta}(s)\big]_{\min_{s'} V_{h+1}^{k,\beta,\eta}(s')+\beta} \Big) \Bigg\|_{(\Lambda_h^k)^{-1}}}_{\text{(iii)}} \\
&\qquad + \lambda \|\phi_i \mathbb{1}_i\|_{(\Lambda_h^k)^{-1}} \cdot \underbrace{\Big\| \mathbb{E}_{s\sim\boldsymbol{\mu}^o} \big[V_{h+1}^{k,\beta,\eta}(s)\big]_{\min_{s'} V_{h+1}^{k,\beta,\eta}(s')+\beta} \Big\|_{(\Lambda_h^k)^{-1}}}_{\text{(iv)}}, \qquad \text{(D.29)}
\end{aligned}
$$

where (D.27) is obtained from the algorithm update formulation and the definition of Λ_h^k, (D.28) is from the definition in (D.23), (D.29) follows from the Cauchy-Schwartz inequality.

For term (iv) in (D.29), since $V_{h+1}^{k,\beta,\eta} \le H$, we have

$$
\begin{aligned}
&\Big\| \mathbb{E}_{s\sim\boldsymbol{\mu}^o} \big[V_{h+1}^{k,\beta,\eta}\big]_{\min_{s'} V_{h+1}^{k,\beta,\eta}(s')+\beta} \Big\|_{(\Lambda_h^k)^{-1}} \\
&\le \lambda_{\max}\big((\Lambda_h^k)^{-1}\big)^{\frac{1}{2}} \Big\| \mathbb{E}_{s\sim\boldsymbol{\mu}^o} \big[V_{h+1}^{k,\beta,\eta}\big]_{\min_{s'} V_{h+1}^{k,\beta,\eta}(s')+\beta} \Big\|_2 \qquad \text{(D.30)} \\
&\le \lambda^{-\frac{1}{2}} H\sqrt{d},
\end{aligned}
$$

where (D.30) is because $\|x\|_A \le \sqrt{\lambda_{\max}(A)}\|x\|_2$.

Before we continue to bound term (iii), we prove a auxiliary result first. That is, $\|\boldsymbol{w}_h^k\|_2 \le H\sqrt{kd/\lambda}$. From the algorithm update formulation, we have

$$
\begin{aligned}
\|\boldsymbol{w}_h^k\|_2 &= \Bigg\| \big(\Lambda_h^k\big)^{-1} \Bigg(\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \big[V_{h+1}^{k,\beta,\eta}(s_{h+1}^\tau)\big]_{\min_{s'} V_{h+1}^{k,\beta,\eta}(s')+\beta} \Bigg) \Bigg\|_2 \\
&\le H \sum_{\tau=1}^{k-1} \Big\| \big(\Lambda_h^k\big)^{-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \Big\|_2 \\
&= H \sum_{\tau=1}^{k-1} \sqrt{\boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top \big(\Lambda_h^k\big)^{-1/2} \big(\Lambda_h^k\big)^{-1} \big(\Lambda_h^k\big)^{-1/2} \boldsymbol{\phi}(s_h^\tau, a_h^\tau)}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{H}{\sqrt{\lambda}} \sum_{\tau=1}^{k-1} \sqrt{\boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top \big(\Lambda_h^k\big)^{-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau)} \\
&\leq \frac{H\sqrt{k}}{\sqrt{\lambda}} \sqrt{\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top \big(\Lambda_h^k\big)^{-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau)} \\
&= \frac{H\sqrt{k}}{\sqrt{\lambda}} \sqrt{\mathrm{tr}\big(\big(\Lambda_h^k\big)^{-1}\big(\Lambda_h^k - \lambda I\big)\big)} \\
&\leq \frac{H\sqrt{k}}{\sqrt{\lambda}} \sqrt{\mathrm{tr}(I)} = H\sqrt{kd/\lambda}.
\end{aligned}
$$

Now we are ready to bound term (iii) in (D.29). Let $\mathcal{V}_h$ denote a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with the following from

$$
V(\cdot) = \min\left\{ \max_{a\in\mathcal{A}} \left\{ \boldsymbol{\phi}(\cdot, a)^\top \boldsymbol{w} + \beta \sum_{i=1}^{d} \|\phi_i(\cdot, a)\mathbb{1}_i\|_{\Lambda^{-1}} \right\}, H - h + 1 \right\},
$$

where the parameters (w, β, Λ) satisfy $\|w\| \leq L$, $\beta \in [0, B]$, $\lambda_{\min} \geq \lambda$. Here we set $L = H\sqrt{kd/\lambda}$ and $B = Hd \cdot \xi_{\mathrm{TV}}$, where $\xi_{\mathrm{TV}} = 720 + 3\sqrt{40 \log\big(96K^{13/2}H|\mathcal{A}|^3/\delta\big)}$.

Let $\mathcal{N}_h(\epsilon; \mathcal{V}_h)$ be the minimum ϵ-cover of $\mathcal{V}_h$, $\mathcal{N}_h(\epsilon; [0, H])$ be the minimum ϵ-cover of $[0, H]$. Therefore, for any $V \in \mathcal{V}_h$ and $\alpha \in [0, H]$, there exists a function $V_\epsilon \in \mathcal{N}_h(\epsilon; \mathcal{V}_h)$ and $\alpha_\epsilon \in \mathcal{N}_h(\epsilon; [0, H])$ such that

$$
\sup_{s\in\mathcal{S}} \big|V(s) - V_\epsilon(s)\big| \leq \epsilon, \ \Big| \min_{s'} V_{h+1}^{k,\beta,\eta}(s') + \beta - \alpha_\epsilon \Big| \leq \epsilon.
$$

Then we have

$$
\begin{aligned}
(\mathrm{iii})^2 &= \bigg\| \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau\Big(\big[V_{h+1}^{k,\beta,\eta}\big]_{\min_{s'} V_{h+1}^{k,\beta,\eta}(s')+\beta} \Big) \bigg\|_{(\Lambda_h^k)^{-1}}^2 \\
&\leq 2\bigg\| \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau\big(\big[V_{h+1}^{k,\beta,\eta}\big]_{\alpha_\epsilon} \big) \bigg\|_{(\Lambda_h^k)^{-1}}^2 \\
&\quad + 2\bigg\| \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau\Big(\big[V_{h+1}^{k,\beta,\eta}\big]_{\min_{s'} V_{h+1}^{k,\beta,\eta}(s')+\beta} - \big[V_{h+1}^{k,\beta,\eta}\big]_{\alpha_\epsilon} \Big) \bigg\|_{(\Lambda_h^k)^{-1}}^2 \qquad \text{(D.31)} \\
&\leq 4\bigg\| \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau\big(\big[V_\epsilon\big]_{\alpha_\epsilon} \big) \bigg\|_{(\Lambda_h^k)^{-1}}^2 + 4\bigg\| \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau\big(\big[V_{h+1}^{k,\beta,\eta}\big]_{\alpha_\epsilon} - \big[V_\epsilon\big]_{\alpha_\epsilon} \big) \bigg\|_{(\Lambda_h^k)^{-1}}^2 \\
&\quad + 2(2\epsilon)^2 k^2/\lambda \qquad \text{(D.32)} \\
&\leq 4\bigg\| \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau\big(\big[V_\epsilon\big]_{\alpha_\epsilon} \big) \bigg\|_{(\Lambda_h^k)^{-1}}^2 + 4(2\epsilon)^2 k^2/\lambda + 2(2\epsilon)^2 k^2/\lambda \\
&= 4\bigg\| \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau\big(\big[V_\epsilon\big]_{\alpha_\epsilon} \big) \bigg\|_{(\Lambda_h^k)^{-1}}^2 + 24\epsilon^2 k^2/\lambda
\end{aligned}
$$

$$\leq 4H^2\left(d\log(1+k/\lambda)+2\log\frac{KH|\mathcal{N}_h(\epsilon;\mathcal{V}_h)|\cdot|\mathcal{N}_h(\epsilon;[0,H])|}{\delta}\right)+24\epsilon^2k^2/\lambda \tag{D.33}$$

$$\leq 4H^2\Big(d\log(1+k/\lambda)+2d\log(1+4L|\mathcal{A}|/\epsilon)+2d^2\log\left(1+8d^{1/2}B^2|\mathcal{A}|^2/(\lambda\epsilon^2)\right)$$
$$+\ 2\log(3H/\epsilon)+2\log(KH/\delta)\Big)+24\epsilon^2k^2/\lambda \tag{D.34}$$

$$\leq 4H^2\Big(2d\log(k)+4d\log\left(4k^{3/2}|\mathcal{A}|/(d^{1/2})\right)+4d^2\log\left(8k^2B^2|\mathcal{A}|^2/(H^2d^{3/2})\right)$$
$$+\ 2\log(3k/d)+2\log(KH/\delta)\Big)+24H^2d^2 \tag{D.35}$$

$$\leq 16H^2d^2\log\left(96K^{13/2}B^2|\mathcal{A}|^3/(Hd^3\delta)\right)+24H^2d^2,$$

where (D.31) and (D.32) is because $\|a+b\|_A^2\leq 2\|a\|_A^2+2\|b\|_A^2$, (D.33) is from Lemma E.7 together with a union bound over all $(k,h)\in[K]\times[H]$, (D.34) makes use of Lemmas D.2 and E.5, we set $\epsilon=Hd/k$ and $\lambda=1$ and apply the inequality $\log(1+x)\leq 2\log x$ for $x\geq 2$ in (D.35).

Combining everything together, recall that we set $\lambda=1$, we have

$$\text{(i)}+\text{(ii)}\leq\left(\sqrt{d}+\sqrt{16H^2d^2\log\left(96K^{13/2}B^2|\mathcal{A}|^3/(Hd^3\delta)\right)+24H^2d^2}+H\sqrt{d}\right)\cdot\|\phi_i\mathbb{1}_i\|_{(\Lambda_h^k)^{-1}}$$
$$\leq 3Hd\sqrt{40\log\left(96K^{13/2}B^2|\mathcal{A}|^3/(Hd^3\delta)\right)}\cdot\|\phi_i\mathbb{1}_i\|_{(\Lambda_h^k)^{-1}}.$$

With our choice of B, and noting that $x\geq\max\left\{(2B+\sqrt{A})^2,1\right\}$ guarantees $x\geq A+B\log x$, we conclude the proof. □

Lemma D.5. If we set the bonus term to be the same as in Lemma D.4, then for any $(k,s,a)\in[K]\times\mathcal{S}\times\mathcal{A}$, with probability at least $1-\delta$, the sum of estimation errors can be bounded as

$$Q_1^{k,\beta,\eta}(s,a)-Q_1^{\pi^k,\beta,\eta}(s,a)\leq 2\cdot\mathbb{E}_{P^{w,k},\pi^k}\left[\sum_{h=1}^H\Gamma_h^k(s_h^k,a_h^k)\right],$$

where $P_h^{w,k}$ is defined by $P_h^{w,k}(s,a)=\boldsymbol{\phi}(s,a)^\top\cdot\boldsymbol{\mu}_h^{w,k}$.

Proof. From the definition, We have that

$$Q_h^{k,\beta,\eta}-Q_h^{\pi^k,\beta,\eta}$$
$$=\Gamma_h^k+\boldsymbol{\phi}^\top\left(\boldsymbol{\theta}_h^k+\boldsymbol{w}_h^k\right)-\boldsymbol{\phi}^\top\left[\boldsymbol{\theta}_h+\inf_{\boldsymbol{\mu}_h\in\Delta(\mathcal{S})}\left(\mathbb{E}_{s\sim\boldsymbol{\mu}_h}[V_{h+1}^{\pi^k,\beta,\eta}]+\beta\mathrm{TV}(\boldsymbol{\mu}_h\|\boldsymbol{\mu}_h^o)\right)\right]$$
$$=\Gamma_h^k+\boldsymbol{\phi}^\top\left(\boldsymbol{\theta}_h^k-\boldsymbol{\theta}_h\right)+\boldsymbol{\phi}^\top\left[\boldsymbol{w}_h^k-\inf_{\boldsymbol{\mu}_h\in\Delta(\mathcal{S})}\left(\mathbb{E}_{s\sim\boldsymbol{\mu}_h}[V_{h+1}^{k,\beta,\eta}]+\beta\mathrm{TV}(\boldsymbol{\mu}_h\|\boldsymbol{\mu}_h^o)\right)\right]$$
$$+\ \boldsymbol{\phi}^\top\left[\inf_{\boldsymbol{\mu}_h\in\Delta(\mathcal{S})}\left(\mathbb{E}_{s\sim\boldsymbol{\mu}_h}[V_{h+1}^{k,\beta,\eta}]+\beta\mathrm{TV}(\boldsymbol{\mu}_h\|\boldsymbol{\mu}_h^o)\right)-\inf_{\boldsymbol{\mu}_h\in\Delta(\mathcal{S})}\left(\mathbb{E}_{s\sim\boldsymbol{\mu}_h}[V_{h+1}^{\pi^k,\beta,\eta}]+\beta\mathrm{TV}(\boldsymbol{\mu}_h\|\boldsymbol{\mu}_h^o)\right)\right]$$
$$\leq 2\cdot\Gamma_h^k+\boldsymbol{\phi}^\top\left[\inf_{\boldsymbol{\mu}_h\in\Delta(\mathcal{S})}\left(\mathbb{E}_{s\sim\boldsymbol{\mu}_h}[V_{h+1}^{k,\beta,\eta}]+\beta\mathrm{TV}(\boldsymbol{\mu}_h\|\boldsymbol{\mu}_h^o)\right)-\inf_{\boldsymbol{\mu}_h\in\Delta(\mathcal{S})}\left(\mathbb{E}_{s\sim\boldsymbol{\mu}_h}[V_{h+1}^{\pi^k,\beta,\eta}]+\beta\mathrm{TV}(\boldsymbol{\mu}_h\|\boldsymbol{\mu}_h^o)\right)\right] \tag{D.36}$$
$$\leq 2\cdot\Gamma_h^k+\boldsymbol{\phi}^\top\mathbb{E}_{s\sim\boldsymbol{\mu}_h^{w,k}}\left[V_{h+1}^{k,\beta,\eta}-V_{h+1}^{\pi^k,\beta,\eta}\right] \tag{D.37}$$
$$\leq 2\cdot\Gamma_h^k+\mathbb{E}_{s\sim P_h^w}\left[V_{h+1}^{k,\beta,\eta}-V_{h+1}^{\pi^k,\beta,\eta}\right],$$

where (D.36) can be derived from Lemma D.4, and (D.37) uses the definition of the worst-case transition. Apply the above inequality recursively, we have

$$Q_1^{k,\beta,\eta} - Q_1^{\pi^k,\beta,\eta} \le 2 \cdot \mathbb{E}_{P^{w,k},\pi^k}\left[\sum_{h=1}^{H} \Gamma_h^k\right].$$

This finishes the proof. □

Proof of Lemma D.1 in TV Setting. The result directly follows from combining Lemma D.4 and Lemma D.5, along with applying a union bound. □

D.3.2 Proof of Lemma D.1 in KL Setting

Lemma D.6 (Dual formulation)**.** (Tang et al., 2025, Proposition 4.5) For the optimization problem $\inf_{\boldsymbol{\mu}\in\Delta(\mathcal{S})}\left(\mathbb{E}_{s\sim\boldsymbol{\mu}}[V_{h+1}] + \beta \mathrm{KL}(\boldsymbol{\mu}\|\boldsymbol{\mu}^o)\right)$, we have its dual formulation as follows

$$\inf_{\boldsymbol{\mu}\in\Delta(\mathcal{S})}\left(\mathbb{E}_{s\sim\boldsymbol{\mu}}[V_{h+1}] + \beta \mathrm{KL}(\boldsymbol{\mu}\|\boldsymbol{\mu}^o)\right) = -\beta \log \mathbb{E}_{s\sim\boldsymbol{\mu}^o}\left[e^{-V_{h+1}(s)/\beta}\right].$$

Lemma D.7 (Optimism)**.** If we set the bonus term as follows

$$\begin{aligned}
\Gamma_h^k(s,a) &= c_{\mathrm{KL}} \sum_{i=1}^{d} \|\phi_i(s,a)\mathbb{1}_i\|_{(\Lambda_h^k)^{-1}}, \\
&\text{where} \quad c_{\mathrm{KL}} = (1 + 2\beta e^{\beta^{-1}H})Hd \cdot \xi_{\mathrm{KL}}, \\
&\text{where} \quad \xi_{\mathrm{KL}} = 80 + \sqrt{40 \log\left(64\beta K^{11/2} H d^{1/2}(1+2\beta e^{\beta^{-1}H})^2 |\mathcal{A}|^3/\delta\right)},
\end{aligned}$$

then for any policy π and any $(k,h,s,a) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}$, with probability at least $1-\delta$, we have $Q_h^{k,\beta,\eta}(s,a) \ge Q_h^{\pi,\beta,\eta}(s,a)$. Specially, by setting $\pi = \pi^*$, we have $Q_h^{k,\beta,\eta}(s,a) \ge Q_h^{\pi^*,\beta,\eta}(s,a)$.

Proof. We prove this by induction. First, $h = H+1$ holds trivially since $Q_{H+1}^{k,\beta,\eta}(s,a) = 0 = Q_{H+1}^{\pi,\beta,\eta}(s,a)$.

Assume $Q_{h+1}^{k,\beta,\eta}(s,a) \ge Q_{h+1}^{\pi,\beta,\eta}(s,a)$ holds, due to the choice of π_{h+1}^k in (4.3), we have

$$\begin{aligned}
V_{h+1}^{k,\beta,\eta}(s) &= \left\langle Q_{h+1}^{k,\beta,\eta}(s,\cdot), \pi_{h+1}^k(\cdot|s)\right\rangle - \Omega_s^\eta(\pi_{h+1}^k) \\
&\ge \left\langle Q_{h+1}^{k,\beta,\eta}(s,\cdot), \pi_{h+1}(\cdot|s)\right\rangle - \Omega_s^\eta(\pi_{h+1}) \\
&\ge \left\langle Q_{h+1}^{\pi,\beta,\eta}(s,\cdot), \pi_{h+1}(\cdot|s)\right\rangle - \Omega_s^\eta(\pi_{h+1}) = V_{h+1}^{\pi,\beta,\eta}(s).
\end{aligned}$$

Recall that we denote $Q_h^{k,\beta,\eta}$ as the optimistic estimation in k-th episode, that is,

$$Q_h^{k,\beta,\eta}(s,a) \leftarrow \min\left\{\boldsymbol{\phi}(s,a)^\top\left(\boldsymbol{\theta}_h^k + \boldsymbol{w}_h^k\right) + \Gamma_h^k(s,a), H-h+1\right\}. \tag{D.38}$$

If $Q_h^{k,\beta,\eta}(s,a) = H-h+1$, then it follows immediately that

$$Q_h^{k,\beta,\eta}(s,a) = H-h+1 \ge Q_h^{\pi,\beta,\eta}(s,a)$$

by the definition of $Q_h^{\pi,\beta,\eta}(s,a)$. Otherwise, we can infer that

$$
\begin{aligned}
&Q_h^{k,\beta,\eta} - Q_h^{\pi,\beta,\eta} \\
&= \Gamma_h^k + \boldsymbol{\phi}^\top(\boldsymbol{\theta}_h^k + \boldsymbol{w}_h^k) - \boldsymbol{\phi}^\top\Big[\boldsymbol{\theta}_h + \inf_{\mu\in\Delta(\mathcal{S})}\big(\mathbb{E}_{s\sim\mu}[V_{h+1}^{\pi,\beta,\eta}] + \beta\mathrm{KL}(\mu\|\mu^o)\big)\Big] \\
&\geq \Gamma_h^k + \boldsymbol{\phi}^\top(\boldsymbol{\theta}_h^k + \boldsymbol{w}_h^k) - \boldsymbol{\phi}^\top\Big[\boldsymbol{\theta}_h + \inf_{\mu\in\Delta(\mathcal{S})}\big(\mathbb{E}_{s\sim\mu}[V_{h+1}^{k,\beta,\eta}] + \beta\mathrm{KL}(\mu\|\mu^o)\big)\Big] \qquad \text{(D.39)} \\
&= \Gamma_h^k + \Big\langle \boldsymbol{\phi}, \boldsymbol{\theta}_h^k - \beta\log\max\big\{\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}\big[e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big], e^{-H/\beta}\big\} - \boldsymbol{\theta}_h + \beta\log\mathbb{E}_{s\sim\mu^o}\big[e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big]\Big\rangle \qquad \text{(D.40)} \\
&\geq \Gamma_h^k - \Big\langle \boldsymbol{\phi}, \big|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h\big| + \beta e^{\beta^{-1}H}\Big|\max\big\{\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}\big[e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big], e^{-H/\beta}\big\} - \mathbb{E}_{s\sim\mu^o}\big[e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big]\Big|\Big\rangle \\
&\geq \Gamma_h^k - \Big\langle \boldsymbol{\phi}, \big|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h\big| + \beta e^{\beta^{-1}H}\Big|\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}\big[e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big] - \mathbb{E}_{s\sim\mu^o}\big[e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big]\Big|\Big\rangle \\
&= \Gamma_h^k - \sum_{i=1}^d \underbrace{\phi_i\mathbb{1}_i^\top\big|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h\big|}_{\text{(i)}} - \beta e^{\beta^{-1}H}\sum_{i=1}^d \underbrace{\phi_i\mathbb{1}_i^\top\Big|\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}\big[e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big] - \mathbb{E}_{s\sim\mu^o}\big[e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big]\Big|}_{\text{(ii)}}, \qquad \text{(D.41)}
\end{aligned}
$$

where (D.39) follows from the induction assumption, and (D.40) is obtained from the algorithm update formulation and the dual formulation.

For term (i) in (D.41), we have

$$
\begin{aligned}
&\big|\phi_i\mathbb{1}_i^\top(\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h)\big| \\
&= \Bigg|\phi_i\mathbb{1}_i^\top(\Lambda_h^k)^{-1}\Bigg(\sum_{\tau=1}^{k-1}\boldsymbol{\phi}(s_h^\tau, a_h^\tau)\cdot r_h^\tau\Bigg) \\
&\qquad - \phi_i\mathbb{1}_i^\top(\Lambda_h^k)^{-1}\Bigg(\sum_{\tau=1}^{k-1}\boldsymbol{\phi}(s_h^\tau, a_h^\tau)\boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top + \lambda\cdot I\Bigg)\cdot\boldsymbol{\theta}_h\Bigg| \\
&= \lambda\big|\phi_i\mathbb{1}_i^\top(\Lambda_h^k)^{-1}\cdot\boldsymbol{\theta}_h\big| \qquad \text{(D.42)} \\
&\leq \lambda\|\phi_i\mathbb{1}_i\|_{(\Lambda_h^k)^{-1}}\cdot\|\boldsymbol{\theta}_h\|_{(\Lambda_h^k)^{-1}} \qquad \text{(D.43)} \\
&\leq \lambda\|\phi_i\mathbb{1}_i\|_{(\Lambda_h^k)^{-1}}\cdot\lambda_{\max}\big((\Lambda_h^k)^{-1}\big)^{\frac{1}{2}}\|\boldsymbol{\theta}_h\|_2 \qquad \text{(D.44)} \\
&\leq \lambda^{\frac{1}{2}}\sqrt{d}\cdot\|\phi_i\mathbb{1}_i\|_{(\Lambda_h^k)^{-1}},
\end{aligned}
$$

where (D.42) is obtained from the algorithm update formulation and the definition of Λ_h^k, (D.43) follows from the Cauchy-Schwartz inequality, (D.44) is because $\|x\|_A \leq \sqrt{\lambda_{\max}(A)}\|x\|_2$.

For term (ii) in (D.41), we decompose it as follows.

$$
\begin{aligned}
&\Big|\phi_i\mathbb{1}_i^\top\Big(\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}\big[e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big] - \mathbb{E}_{s\sim\boldsymbol{\mu}^o}\big[e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big]\Big)\Big| \\
&= \Bigg|\phi_i\mathbb{1}_i^\top(\Lambda_h^k)^{-1}\Bigg(\sum_{\tau=1}^{k-1}\boldsymbol{\phi}(s_h^\tau, a_h^\tau)\cdot e^{-V_{h+1}^{k,\beta,\eta}(s_{h+1}^\tau)/\beta}\Bigg) \\
&\qquad - \phi_i\mathbb{1}_i^\top(\Lambda_h^k)^{-1}\Bigg(\sum_{\tau=1}^{k-1}\boldsymbol{\phi}(s_h^\tau, a_h^\tau)\boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top + \lambda\cdot I\Bigg)\mathbb{E}_{s\sim\boldsymbol{\mu}^o}\big[e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big]\Bigg| \qquad \text{(D.45)}
\end{aligned}
$$

$$
\begin{aligned}
&= \Bigg| \phi_i \mathbb{1}_i^\top \big(\Lambda_h^k\big)^{-1} \Bigg(\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau\big(e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big) \Bigg) \\
&\qquad + \lambda \phi_i \mathbb{1}_i^\top \big(\Lambda_h^k\big)^{-1} \mathbb{E}_{s\sim\boldsymbol{\mu}^o}\big[e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big] \Bigg| \qquad \text{(D.46)} \\
&\le \Bigg| \phi_i \mathbb{1}_i^\top \big(\Lambda_h^k\big)^{-1} \Bigg(\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau\big(e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big) \Bigg) \Bigg| \\
&\qquad + \lambda \Big| \phi_i \mathbb{1}_i^\top \big(\Lambda_h^k\big)^{-1} \mathbb{E}_{s\sim\boldsymbol{\mu}^o}\big[e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big] \Big| \\
&\le \|\phi_i \mathbb{1}_i\|_{(\Lambda_h^k)^{-1}} \cdot \underbrace{\Bigg\| \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau\big(e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big) \Bigg\|_{(\Lambda_h^k)^{-1}}}_{\text{(iii)}} \\
&\qquad + \lambda \|\phi_i \mathbb{1}_i\|_{(\Lambda_h^k)^{-1}} \cdot \underbrace{\Big\| \mathbb{E}_{s\sim\boldsymbol{\mu}^o}\big[e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big] \Big\|_{(\Lambda_h^k)^{-1}}}_{\text{(iv)}}, \qquad \text{(D.47)}
\end{aligned}
$$

where (D.45) is obtained from the algorithm update formulation and the definition of Λ_h^k, (D.46) is from the definition in (D.23), (D.47) follows from the Cauchy-Schwartz inequality.

For term (iv) in (D.47), since $e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta} \le 1$, we have

$$
\begin{aligned}
&\Big\| \mathbb{E}_{s\sim\boldsymbol{\mu}^o}\big[e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big] \Big\|_{(\Lambda_h^k)^{-1}} \\
&\le \lambda_{\max}\big((\Lambda_h^k)^{-1}\big)^{\frac{1}{2}} \Big\| \mathbb{E}_{s\sim\boldsymbol{\mu}^o}\big[e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big] \Big\|_2 \qquad \text{(D.48)} \\
&\le \lambda^{-\frac{1}{2}} \sqrt{d},
\end{aligned}
$$

where (D.48) is because $\|x\|_A \le \sqrt{\lambda_{\max}(A)}\|x\|_2$.

Before we continue to bound term (iii), we prove a auxiliary result first. That is, $\|\boldsymbol{w}_h^k\|_2 \le \sqrt{kd/\lambda}$. From the algorithm update formulation, we have

$$
\begin{aligned}
\Big\| \widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}\big[e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big] \Big\|_2 &= \Bigg\| \big(\Lambda_h^k\big)^{-1} \Bigg(\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot e^{-V_{h+1}^{k,\beta,\eta}(s_{h+1}^\tau)/\beta} \Bigg) \Bigg\|_2 \\
&\le \sum_{\tau=1}^{k-1} \Big\| \big(\Lambda_h^k\big)^{-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \Big\|_2 \\
&= \sum_{\tau=1}^{k-1} \sqrt{\boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top \big(\Lambda_h^k\big)^{-1/2} \big(\Lambda_h^k\big)^{-1} \big(\Lambda_h^k\big)^{-1/2} \boldsymbol{\phi}(s_h^\tau, a_h^\tau)} \\
&\le \frac{1}{\sqrt{\lambda}} \sum_{\tau=1}^{k-1} \sqrt{\boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top \big(\Lambda_h^k\big)^{-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau)} \\
&\le \frac{\sqrt{k}}{\sqrt{\lambda}} \sqrt{\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top \big(\Lambda_h^k\big)^{-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau)}
\end{aligned}
$$

$$= \frac{\sqrt{k}}{\sqrt{\lambda}} \sqrt{\mathrm{tr}\Big(\big(\Lambda_h^k\big)^{-1} \big(\Lambda_h^k - \lambda I\big) \Big)}$$
$$\leq \frac{\sqrt{k}}{\sqrt{\lambda}} \sqrt{\mathrm{tr}(I)} = \sqrt{kd/\lambda}.$$

Therefore,

$$\begin{aligned}
\|\boldsymbol{w}_h^k\|_2 &= \big\| \beta \log \max\big\{ \widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}\big[e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big], e^{-H/\beta} \big\} \big\|_2 \\
&\leq \beta\sqrt{d} \max\Big\{ \log\Big(\big\| \widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}\big[e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big] \big\|_2 \Big), H/\beta \Big\} \\
&\leq \beta\sqrt{d} \max\Big\{ \log\big(\sqrt{kd/\lambda}\big), H/\beta \Big\} \\
&\leq \beta\sqrt{d}\big(\sqrt{kd/\lambda} + H/\beta\big) \\
&\leq 2\beta H d\sqrt{k}.
\end{aligned}$$

Now we are ready to bound term (iii) in (D.47). Let $\mathcal{V}_h$ denote a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with the following from

$$V(\cdot) = \min\Big\{ \max_{a\in\mathcal{A}} \Big\{ \boldsymbol{\phi}(\cdot,a)^\top \boldsymbol{w} + \beta \sum_{i=1}^d \|\phi_i(\cdot,a)\mathbb{1}_i\|_{\Lambda^{-1}} \Big\}, H-h+1 \Big\},$$

where the parameters (w, β, Λ) satisfy $\|w\| \leq L$, $\beta \in [0, B]$, $\lambda_{\min} \geq \lambda$. Here we set $L = 2\beta H d\sqrt{k}$ and $B = (1+2\beta e^{\beta^{-1}H}) H d \cdot \xi_{\mathrm{KL}}$, where $\xi_{\mathrm{KL}} = 80 + \sqrt{40 \log\big(64\beta K^{11/2} H d^{1/2} (1+2\beta e^{\beta^{-1}H})^2 |\mathcal{A}|^3/\delta\big)}$.

Let $\mathcal{N}_h(\epsilon; \mathcal{V}_h)$ be the minimum ϵ-cover of $\mathcal{V}_h$. Therefore, for any $V \in \mathcal{V}_h$ and $\alpha \in [0, H]$, there exists a function $V_\epsilon \in \mathcal{N}_h(\epsilon; \mathcal{V}_h)$ and $\alpha_\epsilon \in \mathcal{N}_h(\epsilon; [0, H])$ such that

$$\sup_{s\in\mathcal{S}} \big|V(s) - V_\epsilon(s)\big| \leq \epsilon.$$

Then we have

$$\begin{aligned}
(\mathrm{iii})^2 &= \Big\| \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau\big(e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta}\big) \Big\|_{(\Lambda_h^k)^{-1}}^2 \\
&\leq 2 \Big\| \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau\big(e^{-V_\epsilon(s)/\beta}\big) \Big\|_{(\Lambda_h^k)^{-1}}^2 \\
&\quad + 2 \Big\| \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau\big(e^{-V_{h+1}^{k,\beta,\eta}(s)/\beta} - e^{-V_\epsilon(s)/\beta}\big) \Big\|_{(\Lambda_h^k)^{-1}}^2 && \text{(D.49)} \\
&\leq 2 \Big\| \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau\big(e^{-V_\epsilon(s)/\beta}\big) \Big\|_{(\Lambda_h^k)^{-1}}^2 + 2(4\epsilon)^2 k^2/\lambda && \text{(D.50)} \\
&\leq 4\Big(d\log(1+k/\lambda) + 2\log\frac{KH|\mathcal{N}_h(\epsilon;\mathcal{V}_h)|}{\delta} \Big) + 32\epsilon^2 k^2/\lambda && \text{(D.51)} \\
&\leq 4H^2\Big(d\log(1+k/\lambda) + 2d\log(1+4L|\mathcal{A}|/\epsilon) + 2d^2\log\big(1+8d^{1/2}B^2|\mathcal{A}|^2/(\lambda\epsilon^2)\big) \\
&\quad + 2\log(KH/\delta) \Big) + 32\epsilon^2 k^2/\lambda && \text{(D.52)}
\end{aligned}$$

$$
\begin{aligned}
&\leq 4H^2\Big(2d\log(k) + 4d\log\left(8\beta k^{3/2}|\mathcal{A}|\right) + 4d^2\log\left(8k^2B^2|\mathcal{A}|^2/(H^2d^{3/2})\right) \\
&\quad + 2\log(KH/\delta)\Big) + 32H^2d^2 \qquad \text{(D.53)} \\
&\leq 16H^2d^2\log\left(64\beta K^{11/2}B^2|\mathcal{A}|^3/(Hd^{3/2}\delta)\right) + 32H^2d^2,
\end{aligned}
$$

where (D.49) is because $\|a+b\|_A^2 \leq 2\|a\|_A^2 + 2\|b\|_A^2$, (D.50) follows from the inequality $|e^x - e^y| \leq e^{|x-y|} - 1$ for $\max\{x, y\} \leq 0$ and $e^t \leq 1 + 2t$ for $t \in [0, 1]$, (D.51) is from Lemma E.7 together with a union bound over all $(k, h) \in [K] \times [H]$, (D.52) makes use of Lemma D.2, we set $\epsilon = Hd/k$ and $\lambda = 1$ and apply the inequality $\log(1 + x) \leq 2\log x$ for $x \geq 2$ in (D.53).

Combining everything together, recall that we set $\lambda = 1$, we have

$$
\begin{aligned}
\text{(i)} + \beta e^{\beta^{-1}H}\text{(ii)} &\leq \sqrt{d}\cdot\|\phi_i\mathbb{1}_i\|_{(\Lambda_h^k)^{-1}} \\
&\quad + \beta e^{\beta^{-1}H}\left(\sqrt{16H^2d^2\log\left(64\beta K^{11/2}B^2|\mathcal{A}|^3/(Hd^{3/2}\delta)\right) + 32H^2d^2} + \sqrt{d}\right)\cdot\|\phi_i\mathbb{1}_i\|_{(\Lambda_h^k)^{-1}} \\
&\leq (1 + 2\beta e^{\beta^{-1}H})Hd\sqrt{40\log\left(64\beta K^{11/2}B^2|\mathcal{A}|^3/(Hd^{3/2}\delta)\right)}\cdot\|\phi_i\mathbb{1}_i\|_{(\Lambda_h^k)^{-1}}.
\end{aligned}
$$

With our choice of B, and noting that $x \geq \max\left\{(2B + \sqrt{A})^2, 1\right\}$ guarantees $x \geq A + B\log x$, we conclude the proof. □

Lemma D.8. If we set the bonus term to be the same as in Lemma D.7, then for any $(k, s, a) \in [K] \times \mathcal{S} \times \mathcal{A}$, with probability at least $1 - \delta$, the sum of estimation errors can be bounded as

$$
Q_1^{k,\beta,\eta}(s, a) - Q_1^{\pi^k,\beta,\eta}(s, a) \leq 2\cdot\mathbb{E}_{P^{w,k},\pi^k}\left[\sum_{h=1}^{H}\Gamma_h^k(s_h^k, a_h^k)\right],
$$

where $P_h^{w,k}$ is defined by $P_h^{w,k}(s, a) = \boldsymbol{\phi}(s, a)^\top\cdot\boldsymbol{\mu}_h^{w,k}$.

Proof. From the definition, We have that

$$
\begin{aligned}
&Q_h^{k,\beta,\eta} - Q_h^{\pi^k,\beta,\eta} \\
&= \Gamma_h^k + \boldsymbol{\phi}^\top\left(\boldsymbol{\theta}_h^k + \boldsymbol{w}_h^k\right) - \boldsymbol{\phi}^\top\left[\boldsymbol{\theta}_h + \inf_{\boldsymbol{\mu}_h\in\Delta(\mathcal{S})}\left(\mathbb{E}_{s\sim\boldsymbol{\mu}_h}[V_{h+1}^{\pi^k,\beta,\eta}] + \beta\mathrm{KL}(\boldsymbol{\mu}_h\|\boldsymbol{\mu}_h^o)\right)\right] \\
&= \Gamma_h^k + \boldsymbol{\phi}^\top\left(\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h\right) + \boldsymbol{\phi}^\top\left[\boldsymbol{w}_h^k - \inf_{\boldsymbol{\mu}_h\in\Delta(\mathcal{S})}\left(\mathbb{E}_{s\sim\boldsymbol{\mu}_h}[V_{h+1}^{k,\beta,\eta}] + \beta\mathrm{KL}(\boldsymbol{\mu}_h\|\boldsymbol{\mu}_h^o)\right)\right] \\
&\quad + \boldsymbol{\phi}^\top\left[\inf_{\boldsymbol{\mu}_h\in\Delta(\mathcal{S})}\left(\mathbb{E}_{s\sim\boldsymbol{\mu}_h}[V_{h+1}^{k,\beta,\eta}] + \beta\mathrm{KL}(\boldsymbol{\mu}_h\|\boldsymbol{\mu}_h^o)\right) - \inf_{\boldsymbol{\mu}_h\in\Delta(\mathcal{S})}\left(\mathbb{E}_{s\sim\boldsymbol{\mu}_h}[V_{h+1}^{\pi^k,\beta,\eta}] + \beta\mathrm{KL}(\boldsymbol{\mu}_h\|\boldsymbol{\mu}_h^o)\right)\right] \\
&\leq 2\cdot\Gamma_h^k + \boldsymbol{\phi}^\top\left[\inf_{\boldsymbol{\mu}_h\in\Delta(\mathcal{S})}\left(\mathbb{E}_{s\sim\boldsymbol{\mu}_h}[V_{h+1}^{k,\beta,\eta}] + \beta\mathrm{KL}(\boldsymbol{\mu}_h\|\boldsymbol{\mu}_h^o)\right) - \inf_{\boldsymbol{\mu}_h\in\Delta(\mathcal{S})}\left(\mathbb{E}_{s\sim\boldsymbol{\mu}_h}[V_{h+1}^{\pi^k,\beta,\eta}] + \beta\mathrm{KL}(\boldsymbol{\mu}_h\|\boldsymbol{\mu}_h^o)\right)\right] \qquad \text{(D.54)} \\
&\leq 2\cdot\Gamma_h^k + \boldsymbol{\phi}^\top\mathbb{E}_{s\sim\boldsymbol{\mu}_h^{w,k}}\left[V_{h+1}^{k,\beta,\eta} - V_{h+1}^{\pi^k,\beta,\eta}\right] \qquad \text{(D.55)} \\
&\leq 2\cdot\Gamma_h^k + \mathbb{E}_{s\sim P_h^w}\left[V_{h+1}^{k,\beta,\eta} - V_{h+1}^{\pi^k,\beta,\eta}\right],
\end{aligned}
$$

where (D.54) can be derived from Lemma D.7, and (D.55) uses the definition of the worst-case transition. Apply the above inequality recursively, we have

$$
Q_1^{k,\beta,\eta} - Q_1^{\pi^k,\beta,\eta} \leq 2\cdot\mathbb{E}_{P^{w,k},\pi^k}\left[\sum_{h=1}^{H}\Gamma_h^k\right].
$$

This finishes the proof. □

Proof of Lemma D.1 in KL Setting. The result directly follows from combining Lemma D.7 and Lemma D.8, along with applying a union bound. □

D.3.3 Proof of Lemma D.1 in χ^2 Setting

Lemma D.9 (Dual formulation)**.** (Tang et al., 2025, Proposition 4.7) For the optimization problem $\inf_{\boldsymbol{\mu}\in\Delta(\mathcal{S})}\big(\mathbb{E}_{s\sim\boldsymbol{\mu}}[V_{h+1}]+\beta\chi^2(\boldsymbol{\mu}\|\boldsymbol{\mu}^o)\big)$, we have its dual formulation as follows

$$\inf_{\boldsymbol{\mu}\in\Delta(\mathcal{S})}\big(\mathbb{E}_{s\sim\boldsymbol{\mu}}[V_{h+1}]+\beta\chi^2(\boldsymbol{\mu}\|\boldsymbol{\mu}^o)\big)=\sup_{\alpha\in[V_{\min},V_{\max}]}\left\{\mathbb{E}_{s\sim\boldsymbol{\mu}^o}\big[V_{h+1}(s)\big]_\alpha-\frac{1}{4\beta}\mathrm{Var}_{s\sim\boldsymbol{\mu}^o}\big[V_{h+1}(s)\big]_\alpha\right\}.$$

Lemma D.10 (Optimism)**.** If we set the bonus term as follows

$$\begin{aligned}\Gamma_h^k(s,a)&=c_{\chi^2}\sum_{i=1}^d\|\phi_i(s,a)\mathbb{1}_i\|_{(\Lambda_h^k)^{-1}},\\ \text{where}\quad c_{\chi^2}&=\big(1+H/(2\beta)\big)Hd\cdot\xi_{\chi^2},\\ \text{where}\quad \xi_{\chi^2}&=720+3\sqrt{40\log\big(96K^6H^5(1+H/(2\beta))^3|\mathcal{A}|^3/\delta\big)},\end{aligned}$$

then for any policy π and any $(k,h,s,a)\in[K]\times[H]\times\mathcal{S}\times\mathcal{A}$, with probability at least $1-\delta$, we have $Q_h^{k,\beta,\eta}(s,a)\geq Q_h^{\pi,\beta,\eta}(s,a)$. Specially, by setting $\pi=\pi^*$, we have $Q_h^{k,\beta,\eta}(s,a)\geq Q_h^{\pi^*,\beta,\eta}(s,a)$.

Proof. We prove this by induction. First, $h=H+1$ holds trivially since $Q_{H+1}^{k,\beta,\eta}(s,a)=0=Q_{H+1}^{\pi,\beta,\eta}(s,a)$.

Assume $Q_{h+1}^{k,\beta,\eta}(s,a)\geq Q_{h+1}^{\pi,\beta,\eta}(s,a)$ holds, due to the choice of π_{h+1}^k in (4.3), we have

$$\begin{aligned}V_{h+1}^{k,\beta,\eta}(s)&=\big\langle Q_{h+1}^{k,\beta,\eta}(s,\cdot),\pi_{h+1}^k(\cdot|s)\big\rangle-\Omega_s^\eta(\pi_{h+1}^k)\\ &\geq\big\langle Q_{h+1}^{k,\beta,\eta}(s,\cdot),\pi_{h+1}(\cdot|s)\big\rangle-\Omega_s^\eta(\pi_{h+1})\\ &\geq\big\langle Q_{h+1}^{\pi,\beta,\eta}(s,\cdot),\pi_{h+1}(\cdot|s)\big\rangle-\Omega_s^\eta(\pi_{h+1})=V_{h+1}^{\pi,\beta,\eta}(s).\end{aligned}$$

Recall that we denote $Q_h^{k,\beta,\eta}$ as the optimistic estimation in k-th episode, that is,

$$Q_h^{k,\beta,\eta}(s,a)\leftarrow\min\big\{\boldsymbol{\phi}(s,a)^\top\big(\boldsymbol{\theta}_h^k+\boldsymbol{w}_h^k\big)+\Gamma_h^k(s,a),H-h+1\big\}.\tag{D.56}$$

If $Q_h^{k,\beta,\eta}(s,a)=H-h+1$, then it follows immediately that

$$Q_h^{k,\beta,\eta}(s,a)=H-h+1\geq Q_h^{\pi,\beta,\eta}(s,a)$$

by the definition of $Q_h^{\pi,\beta,\eta}(s,a)$. Otherwise, we introduce a few auxiliary notations

$$\widetilde{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}=\operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d}\sum_{\tau=1}^k\big([V_{h+1}^{k,\beta,\eta}(s_{h+1}^\tau)]_{\alpha_i}-\boldsymbol{\phi}(s_h^\tau,a_h^\tau)^\top\boldsymbol{w}\big)^2+\lambda\|\boldsymbol{w}\|_2^2,$$

$$\widetilde{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}^2 = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \sum_{\tau=1}^{k} \big([V_{h+1}^{k,\beta,\eta}(s_{h+1}^\tau)]_{\alpha_i}^2 - \boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top \boldsymbol{w}\big)^2 + \lambda\|\boldsymbol{w}\|_2^2,$$

and denote

$$\alpha_i = \operatorname*{argmax}_{\alpha\in[0,H]} \bigg\{ \mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_\alpha + \frac{1}{4\beta}\big(\mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_\alpha\big)^2 - \frac{1}{4\beta}\mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_\alpha^2 \bigg\}.$$

Then

$$\begin{aligned}
&Q_h^{k,\beta,\eta} - Q_h^{\pi,\beta,\eta} \\
&= \Gamma_h^k + \boldsymbol{\phi}^\top\big(\boldsymbol{\theta}_h^k + \boldsymbol{w}_h^k\big) - \boldsymbol{\phi}^\top\Big[\boldsymbol{\theta}_h + \inf_{\mu\in\Delta(\mathcal{S})}\big(\mathbb{E}_{s\sim\mu}[V_{h+1}^{\pi,\beta,\eta}] + \beta\chi^2(\mu\|\mu^o)\big)\Big] \\
&\geq \Gamma_h^k + \boldsymbol{\phi}^\top\big(\boldsymbol{\theta}_h^k + \boldsymbol{w}_h^k\big) - \boldsymbol{\phi}^\top\Big[\boldsymbol{\theta}_h + \inf_{\mu\in\Delta(\mathcal{S})}\big(\mathbb{E}_{s\sim\mu}[V_{h+1}^{k,\beta,\eta}] + \beta\chi^2(\mu\|\mu^o)\big)\Big] \qquad \text{(D.57)} \\
&= \Gamma_h^k + \bigg\langle \boldsymbol{\phi}, \boldsymbol{\theta}_h^k + \sup_{\alpha\in[0,H]}\bigg\{\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_\alpha + \frac{1}{4\beta}\big(\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_\alpha\big)^2 - \frac{1}{4\beta}\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_\alpha^2\bigg\} \\
&\qquad - \boldsymbol{\theta}_h - \sup_{\alpha\in[0,H]}\bigg\{\mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_\alpha + \frac{1}{4\beta}\big(\mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_\alpha\big)^2 - \frac{1}{4\beta}\mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_\alpha^2\bigg\}\bigg\rangle \qquad \text{(D.58)} \\
&\geq \Gamma_h^k + \bigg\langle \boldsymbol{\phi}, \boldsymbol{\theta}_h^k + \bigg\{\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i} + \frac{1}{4\beta}\big(\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}\big)^2 - \frac{1}{4\beta}\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}^2\bigg\} \\
&\qquad - \boldsymbol{\theta}_h - \bigg\{\mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i} + \frac{1}{4\beta}\big(\mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}\big)^2 - \frac{1}{4\beta}\mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}^2\bigg\}\bigg\rangle \qquad \text{(D.59)} \\
&\geq \Gamma_h^k - \bigg\langle \boldsymbol{\phi}, \big|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h\big| + \bigg|\frac{1}{4\beta}\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}^2 - \frac{1}{4\beta}\mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}^2\bigg| \\
&\qquad + \bigg|\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i} + \frac{1}{4\beta}\big(\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}\big)^2 - \mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i} - \frac{1}{4\beta}\big(\mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}\big)^2\bigg|\bigg\rangle \\
&= \Gamma_h^k - \sum_{i=1}^d \phi_i \mathbb{1}_i^\top\big|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h\big| - \frac{1}{4\beta}\sum_{i=1}^d \phi_i \mathbb{1}_i^\top\Big|\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}^2 - \mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}^2\Big| \\
&\qquad - \sum_{i=1}^d \phi_i \mathbb{1}_i^\top\Big|\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i} - \mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}\Big| \cdot \bigg|\frac{1}{4\beta}\Big(\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i} + \mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}\Big) + 1\bigg| \\
&= \Gamma_h^k - \sum_{i=1}^d \phi_i \mathbb{1}_i^\top\big|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h\big| \;-\; \frac{1}{4\beta}\sum_{i=1}^d \phi_i \mathbb{1}_i^\top\Big|\widetilde{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}^2 - \mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}^2\Big| \\
&\qquad \times \underbrace{\bigg|\frac{\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}^2 - \mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}^2}{\widetilde{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}^2 - \mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}^2}\bigg|}_{\text{(i)}} - \sum_{i=1}^d \phi_i \mathbb{1}_i^\top\Big|\widetilde{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i} - \mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}\Big| \\
&\qquad \times \underbrace{\bigg|\bigg[\frac{1}{4\beta}\Big(\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i} + \mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}\Big) + 1\bigg] \cdot \frac{\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i} - \mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}}{\widetilde{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i} - \mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}}\bigg|}_{\text{(ii)}}
\end{aligned} \qquad \text{(D.60)}$$

$$\geq \Gamma_h^k - \sum_{i=1}^{d} \underbrace{\phi_i \mathbb{1}_i^\top \left|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h\right|}_{\text{(iii)}} - \frac{1}{4\beta} \sum_{i=1}^{d} \underbrace{\phi_i \mathbb{1}_i^\top \left|\widetilde{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}^2 - \mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}^2\right|}_{\text{(v)}}$$
$$- \left(1 + \frac{H}{2\beta}\right) \sum_{i=1}^{d} \underbrace{\phi_i \mathbb{1}_i^\top \left|\widetilde{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i} - \mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}\right|}_{\text{(iv)}}, \tag{D.61}$$

where (D.57) follows from the induction assumption, (D.58) is obtained from the algorithm update formulation and the dual formulation, and (D.59) comes from our choice of α_i. To establish in (D.60) that (i) ≤ 1 and (ii) $\leq 1 + H/(2\beta)$ and hence that (D.61) holds, we analyze the value of $\widetilde{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}$ or $\widetilde{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}^2$ under three cases:

$$\begin{cases} \text{(i)} \leq 1 & \widetilde{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}^2 < 0 \\ \text{(i)} \leq 1 & \widetilde{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}^2 \in [0, H] \\ \text{(i)} \leq 1 & \widetilde{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}^2 > H, \end{cases} \qquad \begin{cases} \text{(ii)} \leq 1 + H/(4\beta) & \widetilde{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i} < 0 \\ \text{(ii)} \leq 1 + H/(2\beta) & \widetilde{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i} \in [0, H] \\ \text{(ii)} \leq 1 + H/(2\beta) & \widetilde{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i} > H. \end{cases}$$

For term (iii) in (D.61), we have

$$\begin{aligned} &\left|\phi_i \mathbb{1}_i^\top \left(\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h\right)\right| \\ &= \left| \phi_i \mathbb{1}_i^\top \left(\Lambda_h^k\right)^{-1} \left(\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot r_h^\tau \right) \right. \\ &\qquad \left. - \phi_i \mathbb{1}_i^\top \left(\Lambda_h^k\right)^{-1} \left(\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top + \lambda \cdot I \right) \cdot \boldsymbol{\theta}_h \right| \\ &= \lambda \left|\phi_i \mathbb{1}_i^\top \left(\Lambda_h^k\right)^{-1} \cdot \boldsymbol{\theta}_h\right| && \text{(D.62)} \\ &\leq \lambda \|\phi_i \mathbb{1}_i\|_{(\Lambda_h^k)^{-1}} \cdot \|\boldsymbol{\theta}_h\|_{(\Lambda_h^k)^{-1}} && \text{(D.63)} \\ &\leq \lambda \|\phi_i \mathbb{1}_i\|_{(\Lambda_h^k)^{-1}} \cdot \lambda_{\max}\left(\left(\Lambda_h^k\right)^{-1}\right)^{\frac{1}{2}} \|\boldsymbol{\theta}_h\|_2 && \text{(D.64)} \\ &\leq \lambda^{\frac{1}{2}} \sqrt{d} \cdot \|\phi_i \mathbb{1}_i\|_{(\Lambda_h^k)^{-1}}, \end{aligned}$$

where (D.62) is obtained from the algorithm update formulation and the definition of Λ_h^k, (D.63) follows from the Cauchy-Schwartz inequality, (D.64) is because $\|x\|_A \leq \sqrt{\lambda_{\max}(A)}\|x\|_2$.

For term (iv) in (D.61), we decompose it as follows.

$$\begin{aligned} &\left|\phi_i \mathbb{1}_i^\top \left(\widetilde{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i} - \mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}\right)\right| \\ &= \left| \phi_i \mathbb{1}_i^\top \left(\Lambda_h^k\right)^{-1} \left(\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot [V_{h+1}^{k,\beta,\eta}(s_{h+1}^\tau)]_{\alpha_i} \right) \right. \\ &\qquad \left. - \phi_i \mathbb{1}_i^\top \left(\Lambda_h^k\right)^{-1} \left(\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top + \lambda \cdot I \right) \cdot \mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}(s)]_{\alpha_i} \right| && \text{(D.65)} \\ &= \left| \phi_i \mathbb{1}_i^\top \left(\Lambda_h^k\right)^{-1} \left(\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau\left([V_{h+1}^{k,\beta,\eta}(s)]_{\alpha_i}\right) \right) \right. \end{aligned}$$

$$+ \lambda\phi_i \mathbb{1}_i^\top \big(\Lambda_h^k\big)^{-1} \mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}(s)]_{\alpha_i}\Big| \tag{D.66}$$

$$\leq \Big|\phi_i \mathbb{1}_i^\top \big(\Lambda_h^k\big)^{-1} \bigg(\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau\big([V_{h+1}^{k,\beta,\eta}(s)]_{\alpha_i}\big)\bigg)\Big|$$
$$+ \lambda\Big|\phi_i \mathbb{1}_i^\top \big(\Lambda_h^k\big)^{-1} \mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}(s)]_{\alpha_i}\Big|$$
$$\leq \|\phi_i \mathbb{1}_i\|_{(\Lambda_h^k)^{-1}} \cdot \underbrace{\bigg\|\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau\big([V_{h+1}^{k,\beta,\eta}(s)]_{\alpha_i}\big)\bigg\|_{(\Lambda_h^k)^{-1}}}_{\text{(vi)}}$$
$$+ \lambda\|\phi_i \mathbb{1}_i\|_{(\Lambda_h^k)^{-1}} \cdot \underbrace{\Big\|\mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}(s)]_{\alpha_i}\Big\|_{(\Lambda_h^k)^{-1}}}_{\text{(vii)}}, \tag{D.67}$$

where (D.65) is obtained from the algorithm update formulation and the definition of Λ_h^k, (D.66) is from the definition in (D.23), (D.67) follows from the Cauchy-Schwartz inequality.

For term (v) in (D.61), we decompose it as follows.

$$\Big|\phi_i \mathbb{1}_i^\top \Big(\widetilde{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}^2 - \mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}^2\Big)\Big|$$
$$= \Big|\phi_i \mathbb{1}_i^\top \big(\Lambda_h^k\big)^{-1} \bigg(\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot [V_{h+1}^{k,\beta,\eta}(s_{h+1}^\tau)]_{\alpha_i}^2\bigg)$$
$$- \phi_i \mathbb{1}_i^\top \big(\Lambda_h^k\big)^{-1} \bigg(\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau)\boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top + \lambda \cdot I\bigg) \cdot \mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}(s)]_{\alpha_i}^2\Big| \tag{D.68}$$
$$= \Big|\phi_i \mathbb{1}_i^\top \big(\Lambda_h^k\big)^{-1} \bigg(\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau\big([V_{h+1}^{k,\beta,\eta}(s)]_{\alpha_i}^2\big)\bigg)$$
$$+ \lambda\phi_i \mathbb{1}_i^\top \big(\Lambda_h^k\big)^{-1} \mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}(s)]_{\alpha_i}^2\Big| \tag{D.69}$$
$$\leq \Big|\phi_i \mathbb{1}_i^\top \big(\Lambda_h^k\big)^{-1} \bigg(\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau\big([V_{h+1}^{k,\beta,\eta}(s)]_{\alpha_i}^2\big)\bigg)\Big|$$
$$+ \lambda\Big|\phi_i \mathbb{1}_i^\top \big(\Lambda_h^k\big)^{-1} \mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}(s)]_{\alpha_i}^2\Big|$$
$$\leq \|\phi_i \mathbb{1}_i\|_{(\Lambda_h^k)^{-1}} \cdot \underbrace{\bigg\|\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot \mathrm{err}_h^\tau\big([V_{h+1}^{k,\beta,\eta}(s)]_{\alpha_i}^2\big)\bigg\|_{(\Lambda_h^k)^{-1}}}_{\text{(viii)}}$$
$$+ \lambda\|\phi_i \mathbb{1}_i\|_{(\Lambda_h^k)^{-1}} \cdot \underbrace{\Big\|\mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}(s)]_{\alpha_i}^2\Big\|_{(\Lambda_h^k)^{-1}}}_{\text{(ix)}}, \tag{D.70}$$

where (D.68) is obtained from the algorithm update formulation and the definition of Λ_h^k, (D.69) is from the definition in (D.23), (D.70) follows from the Cauchy-Schwartz inequality.

For term (vii) in (D.67) and For term (ix) in (D.70), since $V_{h+1}^{k,\beta,\eta} \le H$, we have

$$\Big\|\mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}\Big\|_{(\Lambda_h^k)^{-1}} \le \lambda_{\max}\big((\Lambda_h^k)^{-1}\big)^{\frac{1}{2}}\Big\|\mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_{\alpha_i}\Big\|_2 \le \lambda^{-\frac{1}{2}}H\sqrt{d},$$
$$\Big\|\mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]^2_{\alpha_i}\Big\|_{(\Lambda_h^k)^{-1}} \le \lambda_{\max}\big((\Lambda_h^k)^{-1}\big)^{\frac{1}{2}}\Big\|\mathbb{E}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]^2_{\alpha_i}\Big\|_2 \le \lambda^{-\frac{1}{2}}H^2\sqrt{d}.$$

Before we continue to bound term (vi) and (viii), we prove a auxiliary result first. That is, $\|\boldsymbol{w}_h^k\|_2 \le \big(H + H^2/(2\beta)\big)\sqrt{d}$. From the algorithm update formulation, we have

$$\begin{aligned}\|\boldsymbol{w}_h^k\|_2 &= \Big\|\sup_{\alpha\in[0,H]}\Big\{\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_\alpha + \frac{1}{4\beta}\big(\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]_\alpha\big)^2 - \frac{1}{4\beta}\widehat{\mathbb{E}}_{s\sim\boldsymbol{\mu}^o}[V_{h+1}^{k,\beta,\eta}]^2_\alpha\Big\}\Big\|_2 \\ &\le \Big\|\mathbb{1}\Big(H + \frac{H^2}{2\beta}\Big)\Big\|_2 = \Big(H + \frac{H^2}{2\beta}\Big)\sqrt{d}.\end{aligned}$$

Now we are ready to bound term (vi) and (viii). Let $\mathcal{V}_h$ denote a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with the following from

$$V(\cdot) = \min\Big\{\max_{a\in\mathcal{A}}\Big\{\boldsymbol{\phi}(\cdot,a)^\top\boldsymbol{w} + \beta\sum_{i=1}^d\|\phi_i(\cdot,a)\mathbb{1}_i\|_{\Lambda^{-1}}\Big\}, H-h+1\Big\},$$

where the parameters (w,β,Λ) satisfy $\|w\| \le L$, $\beta\in[0,B]$, $\lambda_{\min}\ge\lambda$. Here we set $L = \big(H + H^2/(2\beta)\big)\sqrt{d}$ and $B = \big(1+H/(2\beta)\big)Hd\cdot\xi_{\chi^2}$, where $\xi_{\chi^2} = 720+3\sqrt{40\log\big(96K^6H^5(1+H/(2\beta))^3|\mathcal{A}|^3/\delta\big)}$.

Let $\mathcal{N}_h(\epsilon;\mathcal{V}_h)$ be the minimum ϵ-cover of $\mathcal{V}_h$, $\mathcal{N}_h(\epsilon;[0,H])$ be the minimum ϵ-cover of $[0,H]$. Therefore, for any $V\in\mathcal{V}_h$ and $\alpha\in[0,H]$, there exists a function $V_\epsilon\in\mathcal{N}_h(\epsilon;\mathcal{V}_h)$ and $\alpha_\epsilon\in\mathcal{N}_h(\epsilon;[0,H])$ such that

$$\sup_{s\in\mathcal{S}}\big|V(s) - V_\epsilon(s)\big| \le \epsilon,\ \Big|\min_{s'}V_{h+1}^{k,\beta,\eta}(s') + \beta - \alpha_\epsilon\Big| \le \epsilon.$$

Then we have

$$\begin{aligned}(\text{vi})^2 &= \Big\|\sum_{\tau=1}^{k-1}\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\cdot\mathrm{err}_h^\tau\big([V_{h+1}^{k,\beta,\eta}]_{\alpha_i}\big)\Big\|^2_{(\Lambda_h^k)^{-1}} \\ &\le 2\Big\|\sum_{\tau=1}^{k-1}\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\cdot\mathrm{err}_h^\tau\big([V_{h+1}^{k,\beta,\eta}]_{\alpha_\epsilon}\big)\Big\|^2_{(\Lambda_h^k)^{-1}} \\ &\quad + 2\Big\|\sum_{\tau=1}^{k-1}\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\cdot\mathrm{err}_h^\tau\big([V_{h+1}^{k,\beta,\eta}]_{\alpha_i} - [V_{h+1}^{k,\beta,\eta}]_{\alpha_\epsilon}\big)\Big\|^2_{(\Lambda_h^k)^{-1}} \qquad (\text{D.71}) \\ &\le 4\Big\|\sum_{\tau=1}^{k-1}\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\cdot\mathrm{err}_h^\tau\big([V_\epsilon]_{\alpha_\epsilon}\big)\Big\|^2_{(\Lambda_h^k)^{-1}} + 4\Big\|\sum_{\tau=1}^{k-1}\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\cdot\mathrm{err}_h^\tau\big([V_{h+1}^{k,\beta,\eta}]_{\alpha_\epsilon} - [V_\epsilon]_{\alpha_\epsilon}\big)\Big\|^2_{(\Lambda_h^k)^{-1}} \\ &\quad + 2(2\epsilon)^2k^2/\lambda \qquad (\text{D.72}) \\ &\le 4\Big\|\sum_{\tau=1}^{k-1}\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\cdot\mathrm{err}_h^\tau\big([V_\epsilon]_{\alpha_\epsilon}\big)\Big\|^2_{(\Lambda_h^k)^{-1}} + 4(2\epsilon)^2k^2/\lambda + 2(2\epsilon)^2k^2/\lambda\end{aligned}$$

$$
\begin{aligned}
&\le 4\Big\|\sum_{\tau=1}^{k-1}\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\cdot \mathrm{err}_h^\tau\big([V_\epsilon]_{\alpha_\epsilon}\big)\Big\|^2_{(\Lambda_h^k)^{-1}} + 24\epsilon^2k^2/\lambda \\
&\le 4H^2\bigg(d\log(1+k/\lambda)+2\log\frac{KH|\mathcal{N}_h(\epsilon;\mathcal{V}_h)|\cdot|\mathcal{N}_h(\epsilon;[0,H])|}{\delta}\bigg)+24\epsilon^2k^2/\lambda \qquad \text{(D.73)}\\
&\le 4H^2\Big(d\log(1+k/\lambda)+2d\log(1+4L|\mathcal{A}|/\epsilon)+2d^2\log\big(1+8d^{1/2}B^2|\mathcal{A}|^2/(\lambda\epsilon^2)\big)\\
&\quad +2\log(3H/\epsilon)+2\log(KH/\delta)\Big)+24\epsilon^2k^2/\lambda \qquad \text{(D.74)}\\
&\le 4H^2\Big(2d\log(k)+4d\log\big(4k(1+H/(2\beta))|\mathcal{A}|/(d^{1/2})\big)+4d^2\log\big(8k^2B^2|\mathcal{A}|^2/(H^2d^{3/2})\big)\\
&\quad +2\log(3k/d)+2\log(KH/\delta)\Big)+24H^2d^2 \qquad \text{(D.75)}\\
&\le 16H^2d^2\log\big(96K^6B^2(1+H/(2\beta))|\mathcal{A}|^3/(Hd^3\delta)\big)+24H^2d^2,
\end{aligned}
$$

where (D.71) and (D.72) is because $\|a+b\|_A^2\le 2\|a\|_A^2+2\|b\|_A^2$, (D.73) is from Lemma E.7 together with a union bound over all $(k,h)\in[K]\times[H]$, (D.74) makes use of Lemmas D.2 and E.5, we set $\epsilon=Hd/k$ and $\lambda=1$ and apply the inequality $\log(1+x)\le 2\log x$ for $x\ge 2$ in (D.75).

And

$$
\begin{aligned}
(\text{viii})^2 &= \Big\|\sum_{\tau=1}^{k-1}\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\cdot \mathrm{err}_h^\tau\big([V_{h+1}^{k,\beta,\eta}]^2_{\alpha_i}\big)\Big\|^2_{(\Lambda_h^k)^{-1}}\\
&\le 2\Big\|\sum_{\tau=1}^{k-1}\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\cdot \mathrm{err}_h^\tau\big([V_{h+1}^{k,\beta,\eta}]^2_{\alpha_\epsilon}\big)\Big\|^2_{(\Lambda_h^k)^{-1}}\\
&\quad +2\Big\|\sum_{\tau=1}^{k-1}\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\cdot \mathrm{err}_h^\tau\big([V_{h+1}^{k,\beta,\eta}]^2_{\alpha_i}-[V_{h+1}^{k,\beta,\eta}]^2_{\alpha_\epsilon}\big)\Big\|^2_{(\Lambda_h^k)^{-1}} \qquad \text{(D.76)}\\
&\le 4\Big\|\sum_{\tau=1}^{k-1}\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\cdot \mathrm{err}_h^\tau\big([V_\epsilon]^2_{\alpha_\epsilon}\big)\Big\|^2_{(\Lambda_h^k)^{-1}}+4\Big\|\sum_{\tau=1}^{k-1}\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\cdot \mathrm{err}_h^\tau\big([V_{h+1}^{k,\beta,\eta}]^2_{\alpha_\epsilon}-[V_\epsilon]^2_{\alpha_\epsilon}\big)\Big\|^2_{(\Lambda_h^k)^{-1}}\\
&\quad +2(2H\epsilon)^2k^2/\lambda \qquad \text{(D.77)}\\
&\le 4\Big\|\sum_{\tau=1}^{k-1}\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\cdot \mathrm{err}_h^\tau\big([V_\epsilon]^2_{\alpha_\epsilon}\big)\Big\|^2_{(\Lambda_h^k)^{-1}}+4(2H\epsilon)^2k^2/\lambda+2(2H\epsilon)^2k^2/\lambda\\
&\le 4\Big\|\sum_{\tau=1}^{k-1}\boldsymbol{\phi}(s_h^\tau,a_h^\tau)\cdot \mathrm{err}_h^\tau\big([V_\epsilon]^2_{\alpha_\epsilon}\big)\Big\|^2_{(\Lambda_h^k)^{-1}}+24H^2\epsilon^2k^2/\lambda\\
&\le 4H^2\bigg(d\log(1+k/\lambda)+2\log\frac{KH|\mathcal{N}_h(\epsilon;\mathcal{V}_h)|\cdot|\mathcal{N}_h(\epsilon;[0,H])|}{\delta}\bigg)+24H^2\epsilon^2k^2/\lambda \qquad \text{(D.78)}\\
&\le 4H^2\Big(d\log(1+k/\lambda)+2d\log(1+4L|\mathcal{A}|/\epsilon)+2d^2\log\big(1+8d^{1/2}B^2|\mathcal{A}|^2/(\lambda\epsilon^2)\big)\\
&\quad +2\log(3H/\epsilon)+2\log(KH/\delta)\Big)+24H^2\epsilon^2k^2/\lambda \qquad \text{(D.79)}\\
&\le 4H^2\Big(2d\log(k)+4d\log\big(4kH(1+H/(2\beta))|\mathcal{A}|/(d^{1/2})\big)+4d^2\log\big(8k^2B^2|\mathcal{A}|^2/(d^{3/2})\big)\\
&\quad +2\log(3kH/d)+2\log(KH/\delta)\Big)+24H^2d^2 \qquad \text{(D.80)}
\end{aligned}
$$

$$\leq 16H^2d^2\log\left(96K^6H^3B^2(1+H/(2\beta))|\mathcal{A}|^3/(d^3\delta)\right)+24H^2d^2,$$

where (D.76) and (D.77) is because $\|a+b\|_A^2 \leq 2\|a\|_A^2+2\|b\|_A^2$, (D.78) is from Lemma E.7 together with a union bound over all $(k,h)\in[K]\times[H]$, (D.79) makes use of Lemmas D.2 and E.5, we set $\epsilon = d/k$ and $\lambda = 1$ and apply the inequality $\log(1+x)\leq 2\log x$ for $x\geq 2$ in (D.80).

Combining everything together, recall that we set $\lambda = 1$, we have

$$\begin{aligned}
&\text{(iii)}+\frac{1}{4\beta}\text{(v)}+\left(1+\frac{H}{2\beta}\right)\text{(iv)}\\
&\leq \sqrt{d}\cdot\|\phi_i\mathbb{1}_i\|_{(\Lambda_h^k)^{-1}}\\
&\quad+\frac{1}{4\beta}\left(\sqrt{16H^2d^2\log\left(96K^6H^3B^2(1+H/(2\beta))|\mathcal{A}|^3/(d^3\delta)\right)+24H^2d^2}+H^2\sqrt{d}\right)\cdot\|\phi_i\mathbb{1}_i\|_{(\Lambda_h^k)^{-1}}\\
&\quad+\left(1+\frac{H}{2\beta}\right)\left(\sqrt{16H^2d^2\log\left(96K^6B^2(1+H/(2\beta))|\mathcal{A}|^3/(Hd^3\delta)\right)+24H^2d^2}+H\sqrt{d}\right)\cdot\|\phi_i\mathbb{1}_i\|_{(\Lambda_h^k)^{-1}}\\
&\leq 3\left(1+\frac{H}{2\beta}\right)Hd\sqrt{40\log\left(96K^6H^3B^2(1+H/(2\beta))|\mathcal{A}|^3/(d^3\delta)\right)}\cdot\|\phi_i\mathbb{1}_i\|_{(\Lambda_h^k)^{-1}}.
\end{aligned}$$

With our choice of B, and noting that $x\geq\max\left\{(2B+\sqrt{A})^2,1\right\}$ guarantees $x\geq A+B\log x$, we conclude the proof. □

Lemma D.11. If we set the bonus term to be the same as in Lemma D.10, then for any $(k,s,a)\in[K]\times\mathcal{S}\times\mathcal{A}$, with probability at least $1-\delta$, the sum of estimation errors can be bounded as

$$Q_1^{k,\beta,\eta}(s,a)-Q_1^{\pi^k,\beta,\eta}(s,a)\leq 2\cdot\mathbb{E}_{P^{w,k},\pi^k}\left[\sum_{h=1}^{H}\Gamma_h^k(s_h^k,a_h^k)\right],$$

where $P_h^{w,k}$ is defined by $P_h^{w,k}(s,a)=\boldsymbol{\phi}(s,a)^\top\cdot\boldsymbol{\mu}_h^{w,k}$.

Proof. From the definition, We have that

$$\begin{aligned}
&Q_h^{k,\beta,\eta}-Q_h^{\pi^k,\beta,\eta}\\
&=\Gamma_h^k+\boldsymbol{\phi}^\top\left(\boldsymbol{\theta}_h^k+\boldsymbol{w}_h^k\right)-\boldsymbol{\phi}^\top\left[\boldsymbol{\theta}_h+\inf_{\boldsymbol{\mu}_h\in\Delta(\mathcal{S})}\left(\mathbb{E}_{s\sim\boldsymbol{\mu}_h}[V_{h+1}^{\pi^k,\beta,\eta}]+\beta\chi^2(\boldsymbol{\mu}_h\|\boldsymbol{\mu}_h^o)\right)\right]\\
&=\Gamma_h^k+\boldsymbol{\phi}^\top\left(\boldsymbol{\theta}_h^k-\boldsymbol{\theta}_h\right)+\boldsymbol{\phi}^\top\left[\boldsymbol{w}_h^k-\inf_{\boldsymbol{\mu}_h\in\Delta(\mathcal{S})}\left(\mathbb{E}_{s\sim\boldsymbol{\mu}_h}[V_{h+1}^{k,\beta,\eta}]+\beta\chi^2(\boldsymbol{\mu}_h\|\boldsymbol{\mu}_h^o)\right)\right]\\
&\qquad+\boldsymbol{\phi}^\top\left[\inf_{\boldsymbol{\mu}_h\in\Delta(\mathcal{S})}\left(\mathbb{E}_{s\sim\boldsymbol{\mu}_h}[V_{h+1}^{k,\beta,\eta}]+\beta\chi^2(\boldsymbol{\mu}_h\|\boldsymbol{\mu}_h^o)\right)-\inf_{\boldsymbol{\mu}_h\in\Delta(\mathcal{S})}\left(\mathbb{E}_{s\sim\boldsymbol{\mu}_h}[V_{h+1}^{\pi^k,\beta,\eta}]+\beta\chi^2(\boldsymbol{\mu}_h\|\boldsymbol{\mu}_h^o)\right)\right]\\
&\leq 2\cdot\Gamma_h^k+\boldsymbol{\phi}^\top\left[\inf_{\boldsymbol{\mu}_h\in\Delta(\mathcal{S})}\left(\mathbb{E}_{s\sim\boldsymbol{\mu}_h}[V_{h+1}^{k,\beta,\eta}]+\beta\chi^2(\boldsymbol{\mu}_h\|\boldsymbol{\mu}_h^o)\right)-\inf_{\boldsymbol{\mu}_h\in\Delta(\mathcal{S})}\left(\mathbb{E}_{s\sim\boldsymbol{\mu}_h}[V_{h+1}^{\pi^k,\beta,\eta}]+\beta\chi^2(\boldsymbol{\mu}_h\|\boldsymbol{\mu}_h^o)\right)\right] \qquad \text{(D.81)}\\
&\leq 2\cdot\Gamma_h^k+\boldsymbol{\phi}^\top\mathbb{E}_{s\sim\boldsymbol{\mu}_h^{w,k}}\left[V_{h+1}^{k,\beta,\eta}-V_{h+1}^{\pi^k,\beta,\eta}\right] \qquad \text{(D.82)}\\
&\leq 2\cdot\Gamma_h^k+\mathbb{E}_{s\sim P_h^w}\left[V_{h+1}^{k,\beta,\eta}-V_{h+1}^{\pi^k,\beta,\eta}\right],
\end{aligned}$$

where (D.81) can be derived from Lemma D.10, and (D.82) uses the definition of the worst-case transition. Apply the above inequality recursively, we have

$$Q_1^{k,\beta,\eta} - Q_1^{\pi^k,\beta,\eta} \leq 2 \cdot \mathbb{E}_{P^{w,k},\pi^k}\left[\sum_{h=1}^{H} \Gamma_h^k\right].$$

This finishes the proof. □

Proof of Lemma D.1 in χ^2 Setting. The result directly follows from combining Lemma D.10 and Lemma D.11, along with applying a union bound. □

E Auxiliary Lemmas

Here, we present some auxiliary lemmas which are useful in the proof.

Lemma E.1 (Hoeffding's inequality)**.** (Vershynin, 2018, Theorem 2.2.6) Let $X_1, \cdots, X_T$ be independent random variables. Assume that $X_t \in [0, M]$ for every t with $M > 0$. Let $S_T = \frac{1}{T}\sum_{t=1}^{T} X_t$, then for any $\epsilon > 0$, we have

$$\mathbb{P}\big(\big|S_T - \mathbb{E}[S_T]\big| \geq \epsilon\big) \leq 2\exp\left(-\frac{2T\epsilon^2}{M^2}\right).$$

Lemma E.2 (Self-bounding variance inequality)**.** (Maurer and Pontil, 2009, Theorem 10) Let $X_1, \cdots, X_T$ be independent and identically distributed random variables with finite variance. Assume that $X_t \in [0, M]$ for every t with $M > 0$. Let $S_T^2 = \frac{1}{T}\sum_{t=1}^{T} X_t^2 - (\frac{1}{T}\sum_{t=1}^{T} X_t)^2$, then for any $\epsilon > 0$, we have

$$\mathbb{P}\big(\big|S_T - \sqrt{\mathrm{Var}(X_1)}\big| \geq \epsilon\big) \leq 2\exp\left(-\frac{T\epsilon^2}{2M^2}\right).$$

Lemma E.3. (Weissman et al., 2003, Theorem 2.1) Let P be a probability distribution over $\mathcal{S} = \{s_1, \cdots, s_S\}$, $X_1, \cdots, X_T$ be independent and identically distributed random variables distributed according to P. Let $\widehat{P}(s) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{X_t = s\}$, then for any $\epsilon > 0$, we have

$$\mathbb{P}\big(\big\|P - \widehat{P}\big\|_1 \geq \epsilon\big) \leq 2^S \exp\left(-\frac{T\epsilon^2}{2}\right).$$

Lemma E.4. (Panaganti and Kalathil, 2022, Lemma 7) We define $\mathcal{V} = \big\{V \in \mathbb{R}^S : \|V\|_\infty \leq V_{\max}\big\}$. Let $\mathcal{N}_\mathcal{V}(\epsilon)$ be a minimal ϵ-cover of $\mathcal{V}$ with respect to the distance metric $d(V, V') = \|V - V'\|_\infty$ for some fixed $\epsilon \in (0, 1)$. Then we have

$$\log|\mathcal{N}_\mathcal{V}(\epsilon)| \leq |\mathcal{S}| \cdot \log\left(\frac{3V_{\max}}{\epsilon}\right).$$

Lemma E.5. (Van Handel, 2014, Lemma 5.13) Denote the ϵ-covering number of the closed interval $[a, b]$ for some real number $b > a$ with respect to the distance metric $\mathrm{dist}(\alpha_1, \alpha_2) = |\alpha_1 - \alpha_2|$ as $\mathcal{N}_h(\epsilon; [a, b])$, then we have $\mathcal{N}_h(\epsilon; [a, b]) \leq 3(b - a)/\epsilon$.

Lemma E.6. (Jin et al., 2020, Lemma D.5) For any $\epsilon > 0$, the ϵ-covering number of the Euclidean ball in $\mathbb{R}^d$ with radius $R > 0$ is upper bounded by $(1 + 2R/\epsilon)^d$.

Lemma E.7. (Abbasi-Yadkori et al., 2011, Theorem 1) Let $\{\epsilon_t\}_{t=1}^{\infty}$ be a real-valued stochastic process with corresponding filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$. Let $\epsilon_t|\mathcal{F}_{t-1}$ be mean-zero and σ-sub-Gaussian. Let $\{\boldsymbol{\phi}_t\}_{t=1}^{\infty}$ be a $\mathbb{R}^d$-valued stochastic process where ϕ_t is $\mathcal{F}_{t-1}$ measurable. Assume Λ_0 is a $d \times d$ positive define matrix, and let $\Lambda_t = \Lambda_0 + \sum_{s=1}^{t} \boldsymbol{\phi}_s \boldsymbol{\phi}_s^\top$. Then for any $\delta > 0$, with probability at least $1 - \delta$, we have for all $t \geq 0$,

$$\left\| \sum_{s=1}^{t} \boldsymbol{\phi}_s \epsilon_s \right\|_{\Lambda_t^{-1}}^2 \leq 2\sigma^2 \log \left[\frac{\det(\Lambda_t)^{\frac{1}{2}} \det(\Lambda_0)^{-\frac{1}{2}}}{\delta} \right].$$

Lemma E.8. (Liu and Xu, 2024a, Corollary 5.3) For all $(\pi, h) \in \Pi \times [H]$, assume that

$$\mathbb{E}_\pi[\boldsymbol{\phi}(s_h, a_h)\boldsymbol{\phi}(s_h, a_h)^\top] \geq \alpha I,$$

where $\alpha > 0$. Then with probability at least $1 - \delta$, we have for all $(k, h) \in [K] \times [H]$,

$$\lambda_{\min}(\Lambda_h^k) \geq \max\left\{\alpha(k-1) + \lambda - \sqrt{32k \log(dKH/\delta)}, \lambda\right\}.$$

References

Abbasi-Yadkori, Y., Pál, D. and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems* **24**. 5, 14, 54

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T. et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* . 1

Blanchet, J., Lu, M., Zhang, T. and Zhong, H. (2023). Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *Advances in Neural Information Processing Systems* **36** 66845–66859. 3

Derman, E., Geist, M. and Mannor, S. (2021). Twice regularized mdps and the equivalence between robustness and regularization. *Advances in Neural Information Processing Systems* **34** 22274–22287. 1, 3, 5, 7

Eysenbach, B. and Levine, S. (2021). Maximum entropy rl (provably) solves some robust rl problems. *arXiv preprint arXiv:2103.06257* . 1

Geist, M., Scherrer, B. and Pietquin, O. (2019). A theory of regularized markov decision processes. In *International Conference on Machine Learning*. PMLR. 1

Ghosh, D., Atia, G. K. and Wang, Y. (2025). Provably near-optimal distributionally robust reinforcement learning in online settings. *arXiv preprint arXiv:2508.03768* . 3

He, Y., Liu, Z., Wang, W. and Xu, P. (2025). Sample complexity of distributionally robust off-dynamics reinforcement learning with online interaction. In *Forty-second International Conference on Machine Learning*. 1, 2, 3, 4, 5, 8, 9, 22, 25, 27

Husain, H., Ciosek, K. and Tomioka, R. (2021). Regularized policies are reward robust. In *International Conference on Artificial Intelligence and Statistics*. PMLR. 1

Iyengar, G. N. (2005). Robust dynamic programming. *Mathematics of Operations Research* **30** 257–280. 1, 2

Jin, C., Yang, Z., Wang, Z. and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*. PMLR. 9, 54

Liu, Z., Wang, W. and Xu, P. (2024). Upper and lower bounds for distributionally robust off-dynamics reinforcement learning. *arXiv preprint arXiv:2409.20521* . 2

Liu, Z. and Xu, P. (2024a). Distributionally robust off-dynamics reinforcement learning: Provable efficiency with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*. PMLR. 2, 3, 5, 8, 9, 13, 14, 35, 54

Liu, Z. and Xu, P. (2024b). Minimax optimal and computationally efficient algorithms for distributionally robust offline reinforcement learning. In *Advances in Neural Information Processing Systems*, vol. 37. 3

Liu, Z. and Xu, P. (2025). Linear mixture distributionally robust markov decision processes. *arXiv preprint arXiv:2505.18044* . 3

Lu, M., Zhong, H., Zhang, T. and Blanchet, J. (2024). Distributionally robust reinforcement learning with interactive data collection: Fundamental hardness and near-optimal algorithms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2, 3, 5, 8

Mannor, S., Mebel, O. and Xu, H. (2016). Robust mdps with k-rectangular uncertainty. *Mathematics of Operations Research* **41** 1484–1509. 2

Maurer, A. and Pontil, M. (2009). Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740* . 53

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D. and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*. PmLR. 1

Nilim, A. and El Ghaoui, L. (2005). Robust control of markov decision processes with uncertain transition matrices. *Operations Research* **53** 780–798. 1, 2

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems* **35** 27730–27744. 1

Panaganti, K. and Kalathil, D. (2022). Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*. PMLR. 3, 53

Panaganti, K., Wierman, A. and Mazumdar, E. (2024). Model-free robust ϕ-divergence reinforcement learning using both offline and online data. In *Proceedings of the 41st International Conference on Machine Learning*. 2, 3, 4

Panaganti, K., Xu, Z., Kalathil, D. and Ghavamzadeh, M. (2022). Robust reinforcement learning using offline data. *Advances in Neural Information Processing Systems* **35** 32211–32224. 3

Rolfe, J. T. (2016). Discrete variational autoencoders. *arXiv preprint arXiv:1609.02200* . 14

Schulman, J., Levine, S., Abbeel, P., Jordan, M. and Moritz, P. (2015). Trust region policy optimization. In *International conference on machine learning*. PMLR. 1

Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* . 1

Shi, L. and Chi, Y. (2024). Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *Journal of Machine Learning Research* **25** 1–91. 3

Shi, L., Li, G., Wei, Y., Chen, Y., Geist, M. and Chi, Y. (2024). The curious price of distributional robustness in reinforcement learning with a generative model. *Advances in Neural Information Processing Systems* **36**. 3

Tang, C., Liu, Z. and Xu, P. (2025). Robust offline reinforcement learning with linearly structured f-divergence regularization. In *Forty-second International Conference on Machine Learning*. 3, 4, 9, 36, 41, 46

Van Handel, R. (2014). Probability in high dimension. Tech. rep. 53

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, vol. 47. Cambridge university press. 53

Vieillard, N., Kozuno, T., Scherrer, B., Pietquin, O., Munos, R. and Geist, M. (2020). Leverage the average: an analysis of kl regularization in reinforcement learning. *Advances in Neural Information Processing Systems* **33** 12163–12174. 1

Wang, H., Shi, L. and Chi, Y. (2024a). Sample complexity of offline distributionally robust linear markov decision processes. In *Reinforcement Learning Conference*. 3

Wang, S., Si, N., Blanchet, J. and Zhou, Z. (2023). A finite sample complexity bound for distributionally robust q-learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR. 3

Wang, S., Si, N., Blanchet, J. and Zhou, Z. (2024b). Sample complexity of variance-reduced distributionally robust q-learning. *Journal of Machine Learning Research* **25** 1–77. 3

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S. and Weinberger, M. J. (2003). Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep* 125. 53

Wiesemann, W., Kuhn, D. and Rustem, B. (2013). Robust markov decision processes. *Mathematics of Operations Research* **38** 153–183. 2

Xu, H. and Mannor, S. (2006). The robustness-performance tradeoff in markov decision processes. *Advances in Neural Information Processing Systems* **19**. 2

Yang, W., Li, X. and Zhang, Z. (2019). A regularized approach to sparse optimal policy in reinforcement learning. *Advances in Neural Information Processing Systems* **32**. 1

Yang, W., Wang, H., Kozuno, T., Jordan, S. M. and Zhang, Z. (2023). Robust markov decision processes without model estimation. *arXiv preprint arXiv:2302.01248* . 1, 2, 3, 4

Yang, W., Zhang, L. and Zhang, Z. (2022). Toward theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics* **50** 3223–3248. 3

Zhang, R., Hu, Y. and Li, N. (2024). Soft robust mdps and risk-sensitive mdps: Equivalence, policy gradient, and sample complexity. In *The Twelfth International Conference on Learning Representations*. 1, 2, 3, 4

Zhao, H., Ye, C., Xiong, W., Gu, Q. and Zhang, T. (2025). Logarithmic regret for online kl-regularized reinforcement learning. *arXiv preprint arXiv:2502.07460* . 1

Zhou, Z., Zhou, Z., Bai, Q., Qiu, L., Blanchet, J. and Glynn, P. (2021). Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR. 3