FAST RATES FOR OFFLINE Q-LEARNING WITH KL REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper investigates offline Q-learning for finite-horizon Markov decision processes with Kullback-Leibler (KL) regularization. Recent studies have shown that KL regularization can yield fast rates in offline learning. However, it remains unclear whether Q-learning, a stochastic approximation style algorithm, can achieve such rates in the offline setting. We first propose KL-regularized Q-Learning (KLQ), a pessimism-free algorithm for offline KL-regularized reinforcement learning. By exploiting the quadratic relationship between policy suboptimality and Q-estimation error induced by KL regularization, we show that this quadratic error structure can be preserved through the Q-learning update. Under single-policy concentrability, KLQ achieves a sample complexity of $\widetilde{O}(1/\epsilon)$ to find an ϵ-optimal policy. This sample complexity, however, has an exponential dependence on the regularization parameter and the horizon. To remove this explicit exponential dependence, we further propose KL-regularized Pessimistic Q-Learning (KLPQ), which incorporates a pessimistic bonus and a monotone Q-update. We prove that KLPQ retains the same $\widetilde{O}(1/\epsilon)$ sample complexity while removing this exponential dependence. To the best of our knowledge, these results provide the first finite-sample guarantees for Q-learning that establish a $1/\epsilon$ dependence on the target accuracy ϵ for offline learning in tabular finite-horizon KL-regularized MDPs under single-policy concentrability.

1 INTRODUCTION

Offline reinforcement learning (RL) aims to learn a near-optimal policy from a fixed dataset without further interaction with the environment (Levine et al., 2020). This setting is particularly relevant when additional data collection is costly, risky, or time-consuming, with applications including robotics and healthcare (Levine et al., 2020; Fujimoto et al., 2019). A central challenge in offline RL is the distribution shift induced by limited data coverage. Since the learner cannot collect additional samples, the learned policy may select actions that are poorly represented in the offline dataset, leading to inaccurate value estimates and extrapolation error (Fujimoto et al., 2019; Kumar et al., 2020). One common approach to mitigate this issue is behavior regularization, which constrains the learned policy toward the behavior policy represented in the offline data (Wu et al., 2019). More generally, policy regularization can be imposed with respect to a reference policy. Among these methods, the Kullback–Leibler (KL) divergence is a common choice in offline policy regularization (Wu et al., 2019), and has also become a standard component of reinforcement learning from human feedback and preference based fine-tuning of large language models, where the learned policy is regularized toward a reference model (Ouyang et al., 2022; Rafailov et al., 2023; Xiong et al., 2023).

The widespread use of KL regularization has motivated a growing line of theoretical work on its statistical properties in offline learning. In particular, recent studies have shown that KL regularization can substantially change the statistical behavior of offline learning. In single-step decision problems, fast rates have been established for KL-regularized contextual and multi-armed bandits (Zhao et al., 2025a; 2026b; Ji et al., 2026a). Related results have also been established under forward-KL regularization (Zhao et al., 2026a). With KL regularization, these works establish n^{-1}-type rates, in contrast to the $n^{-1/2}$-type rates typically obtained in standard unregularized settings. Fast rates have also been established for single-step zero-sum games (Zhang et al., 2026; Chen & Zhang, 2026b). More recently, these results have been extended to multi-step sequential decision problems. In par-

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

ticular, Chen et al. (2026) establish an $\widetilde{O}(1/n)$ statistical rate for offline two-player zero-sum Markov games with KL regularization. From an algorithmic perspective, the methods of Chen et al. (2026) can be regarded as fitted Q-iteration. This class of methods is based on the principle of backward induction or value iteration, where Q-functions are estimated stagewise from the end of the horizon to the beginning. In contrast, another important class of algorithms is Q-learning, which adopts incremental temporal difference updates (Watkins & Dayan, 1992). The two classes therefore follow different update schemes. Compared with methods based on backward induction, the stochastic approximation update scheme of Q-learning also introduces quite different challenges in the theoretical analysis. Characterizing the sample efficiency of Q-learning remains a central problem in the statistical theory of reinforcement learning (Li et al., 2024; Liu et al., 2025). In the offline RL setting, Shi et al. (2022) proposed a pessimistic Q-learning algorithm and established finite-sample guarantees under the standard unregularized objective and single-policy concentrability. However, it remains unclear whether Q-learning can attain these fast rates for the KL-regularized objective. This leads to the following question:

Can offline Q-learning achieve fast rates with KL regularization?

The above question poses three main analytical challenges. First, due to its stochastic approximation updates, Q-learning involves complicated statistical dependencies among Q-value estimates across different time steps, raising the question of whether the squared error structure induced by KL regularization can be preserved under such updates. Second, existing finite-sample analyses of offline Q-learning under single-policy concentrability explicitly rely on pessimism, raising the question of whether pessimism is necessary in this setting. Third, under single-policy concentrability, it remains unclear whether incorporating pessimism when learning a KL-regularized objective can lead to a sharper sample-complexity guarantee for Q-learning.

In this paper, we address these challenges and therefore answer this question affirmatively by developing two offline Q-learning algorithms for tabular finite-horizon KL-regularized MDPs. For an MDP with S states and A actions, let $\eta > 0$ denote the inverse temperature and let $C^\star_\eta$ denote the single-policy concentrability coefficient for the corresponding KL-regularized optimal policy. Given K offline trajectories of horizon H, our main contributions are summarized as follows.

- We first introduce *KL-regularized Q-Learning (KLQ)*, a pessimism-free Q-learning algorithm for the KL-regularized objective. Under single-policy concentrability, we prove that KLQ achieves $\widetilde{O}\big(\eta e^{\eta H(H+1)} H^8 S A C^\star_\eta/\epsilon\big)$ sample complexity. This result is new in two respects: in terms of algorithmic paradigms, we establish the first $\widetilde{O}\left(1/\epsilon\right)$ guarantee for Q-learning in tabular finite-horizon KL-regularized MDPs; in terms of offline RL algorithm design, we establish the first fast rate in this setting under single-policy concentrability without explicit pessimism.
- We further introduce *KL-regularized Pessimistic Q-Learning (KLPQ)*, which incorporates a pessimistic bonus and a monotone Q-update. We prove that KLPQ achieves an improved sample complexity of $\widetilde{O}\left(\eta H^7 S A C^\star_\eta/\epsilon\right)$. This result shows that, for Q-learning under single-policy concentrability, the pessimistic construction provides a mechanism for achieving a sharper sample-complexity guarantee.
- Technically, as detailed in Section 5, our analyses are new in two respects. First, by leveraging KL regularization, we establish the key squared-error recursion for Q-learning, which, for the first time, shows that the statistical benefit brought by KL regularization also applies to the stochastic approximation updates of Q-learning, in contrast to prior multi-step fast rate results established for algorithms based on backward induction (Tiapkin et al., 2023; 2024; Zhao et al., 2025b; Chen et al., 2026; Hong et al., 2026). Second, by leveraging pessimism, we sharpen the dependence of policy suboptimality on the squared Q-estimation error by removing the exponential factor.

1.1 Related Works

Offline Reinforcement Learning. Offline RL learns from a fixed dataset without further interaction with the environment, making insufficient coverage a central statistical challenge (Levine et al., 2020). A large body of theoretical work addresses this difficulty under partial coverage conditions, with pessimism playing a particularly important role (Rashidinejad et al., 2021; Xie et al.,

2021; Zhan et al., 2022; Jin et al., 2021). This principle has been incorporated into both model-based methods (Rashidinejad et al., 2021; Xie et al., 2021; Uehara & Sun, 2021) and model-free Q-learning methods (Shi et al., 2022; Yan et al., 2023). Among these works, Shi et al. (2022) is most closely related to ours from an algorithmic perspective: they establish finite-sample guarantees for offline Q-learning in tabular finite-horizon MDPs under single-policy concentrability. In contrast to their unregularized setting, we study Q-learning for a KL-regularized objective and consider both pessimism-free and pessimistic algorithms.

KL-Regularized Learning. KL regularization can fundamentally change the statistical behavior of learning problems. Fast rates have been established for maximum entropy exploration (Tiapkin et al., 2023), sequential RL with demonstration regularization (Tiapkin et al., 2024), and KL-regularized contextual bandits and RLHF (Zhao et al., 2025a). In particular, the entropy regularization considered by Tiapkin et al. (2023) is equivalent to KL regularization with a uniform reference policy, up to an additive constant. In the online setting, Zhao et al. (2025b) establish logarithmic regret for KL-regularized episodic RL, while Hong et al. (2026) study KL-regularized RL with function approximation under misspecification. Logarithmic regret has also been established for KL-regularized multi-armed bandits and zero-sum Markov games (Ji et al., 2026b; Nayak et al., 2025). In offline learning, fast rates are known for KL-regularized contextual and multi-armed bandits (Zhao et al., 2026b; Ji et al., 2026a; Zhao et al., 2026a), as well as preference learning and games (Wu et al., 2025; Zhang et al., 2026; Chen & Zhang, 2026b). The role of pessimism has also been studied in this setting. Ji et al. (2026a) establish matching upper and lower bounds across different regularization regimes for offline KL-regularized multi-armed bandits, while Wu et al. (2025); Zhang et al. (2026); Chen & Zhang (2026b) obtain fast rates without explicit pessimistic estimates in their respective settings. For multi-step sequential problems, Chen et al. (2026) establish a $\widetilde{O}(1/n)$ rate for offline Markov games with KL regularization. Their methods can be regarded as fitted Q-iteration based on backward induction and thus different from ours.

Q-Learning. Q-learning is a classical model-free method that uses stochastic approximation to solve the Bellman optimality equation (Watkins & Dayan, 1992; Even-Dar & Mansour, 2003). Its non-asymptotic theory has been developed under several different settings. Under a generative model, Li et al. (2021a) studied synchronous Q-learning in both discounted and finite-horizon MDPs, while Li et al. (2024) provided a tight sample-complexity characterization for synchronous Q-learning in discounted MDPs, exhibiting an $\widetilde{O}(\epsilon^{-2})$ dependence for ℓ_∞ Q-function estimation. In online RL, Q-learning has been studied for finite-horizon episodic MDPs (Jin et al., 2018; Li et al., 2021b) and in asynchronous settings with Markovian trajectories (Qu & Wierman, 2020; Li et al., 2020). For offline RL, Yan et al. (2023) studied pessimistic asynchronous Q-learning for discounted MDPs with Markovian data under single-policy concentrability, while Shi et al. (2022) established finite-sample guarantees for pessimistic Q-learning in tabular finite-horizon MDPs under single-policy concentrability. Under single-policy concentrability, both offline guarantees exhibit a $\widetilde{O}(\epsilon^{-2})$ dependence on the target policy suboptimality. Different from these works, we consider finite-horizon KL-regularized MDPs and aim to achieve $\widetilde{O}(\epsilon^{-1})$ sample complexity under single-policy concentrability.

1.2 Notation

For a positive integer n, we write $[n] := \{1, \dots, n\}$. For a finite set $\mathcal{X}$, let $\Delta(\mathcal{X})$ denote the probability simplex over $\mathcal{X}$. For functions $f, g : \mathcal{X} \to \mathbb{R}$, we write $\langle f, g \rangle_{\mathcal{X}} := \sum_{x \in \mathcal{X}} f(x)g(x)$, and omit the subscript when the underlying set is clear from context. For $p \in \Delta(\mathcal{X})$, we denote its support by $\mathrm{supp}(p) := \{x \in \mathcal{X} : p(x) > 0\}$. For a function $f : \mathcal{X} \to \mathbb{R}$ and $p \in \Delta(\mathcal{X})$, we use $\mathbb{E}_{x \sim p}[f(x)] := \sum_{x \in \mathcal{X}} p(x) f(x)$ when $\mathcal{X}$ is finite. We use $\mathbf{1}\{\mathcal{E}\}$ for the indicator of an event $\mathcal{E}$, and write $x \vee y := \max\{x, y\}$ for scalars. Inequalities between vectors or functions are understood pointwise unless otherwise stated. We use $\widetilde{O}(\cdot)$ to suppress logarithmic factors in the relevant problem parameters.

2 PROBLEM SETUP

2.1 FINITE-HORIZON KL-REGULARIZED MDPS

Markov Decision Process. We consider a finite-horizon episodic Markov decision process (MDP) denoted by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \rho, \mathbb{P}, r)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action sets, H is the episode length, and $\rho \in \Delta(\mathcal{S})$ is the initial state distribution. We write $S := |\mathcal{S}|$, $A := |\mathcal{A}|$, and $L_h := H - h + 1$ with $L_{H+1} := 0$. The transition kernel is $\mathbb{P} := \{\mathbb{P}_h\}_{h=1}^H$, where $\mathbb{P}_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, and the deterministic reward function is $r := \{r_h\}_{h=1}^H$, where $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$. We write $\mathbb{P}_{h,s,a} := \mathbb{P}_h(\cdot \mid s, a)$. For $f : \mathcal{S} \to \mathbb{R}$, the transition operator is $(\mathbb{P}_h f)(s, a) := \langle \mathbb{P}_{h,s,a}, f \rangle_{\mathcal{S}} = \sum_{s' \in \mathcal{S}} \mathbb{P}_h(s' \mid s, a) f(s')$. A Markov policy is $\pi := \{\pi_h\}_{h=1}^H$ with $\pi_h(\cdot \mid s) \in \Delta(\mathcal{A})$. Starting from $s_1 \sim \rho$, a trajectory is generated according to $a_h \sim \pi_h(\cdot \mid s_h)$ and $s_{h+1} \sim \mathbb{P}_h(\cdot \mid s_h, a_h)$. We denote the state occupancy by $d_h^\pi(s) := \Pr_{\rho,\pi,\mathbb{P}}(s_h = s)$ for $h \in [H + 1]$, and the state-action occupancy by $d_h^\pi(s, a) := d_h^\pi(s)\pi_h(a \mid s)$ for $h \in [H]$. Accordingly, $d_1^\pi(s) = \rho(s)$ and $d_{h+1}^\pi(s') = \sum_{s,a} d_h^\pi(s, a)\mathbb{P}_h(s' \mid s, a)$ for $h \in [H]$.

KL-regularized Value Functions. We now incorporate KL regularization into the MDP defined above. Following the standard framework of regularized MDPs (Geist et al., 2019), fix a reference policy $\pi^{\mathrm{ref}} := \{\pi_h^{\mathrm{ref}}\}_{h=1}^H$ and an inverse temperature $\eta > 0$. The coefficient $1/\eta$ determines the strength of the KL regularization with respect to the reference policy. For any policy π, we define its KL-regularized value function at step h by

$$V_h^\pi(s) := \mathbb{E}_{\pi,\mathbb{P}}\left[\sum_{t=h}^{H}\left(r_t(s_t, a_t) - \frac{1}{\eta}\log\frac{\pi_t(a_t \mid s_t)}{\pi_t^{\mathrm{ref}}(a_t \mid s_t)}\right)\middle| s_h = s\right].$$

Set $V_{H+1}^\pi \equiv 0$ and define the corresponding Q-function by

$$Q_h^\pi(s, a) := r_h(s, a) + (\mathbb{P}_h V_{h+1}^\pi)(s, a).$$

We use the standard convention that $\mathrm{KL}(p||q) = +\infty$ whenever p is not absolutely continuous with respect to q, and otherwise define $\mathrm{KL}(p||q) := \sum_a p(a)\log\frac{p(a)}{q(a)}$, with $0\log 0 := 0$ and $0\log(0/0) := 0$. Under this convention, the corresponding value functions $V_h^\pi(s)$ and $Q_h^\pi(s, a)$ may take the value $-\infty$. Throughout the paper, we adopt the convention $0 \cdot (\pm\infty) = 0$. These functions satisfy the regularized Bellman consistency identity

$$V_h^\pi(s) = \langle \pi_h(\cdot \mid s), Q_h^\pi(s, \cdot)\rangle_{\mathcal{A}} - \frac{1}{\eta}\mathrm{KL}\left(\pi_h(\cdot \mid s)\middle\|\pi_h^{\mathrm{ref}}(\cdot \mid s)\right).$$

Regularized Bellman Optimality. For $Q_h(s, \cdot) \in \mathbb{R}^A$, define the regularized Bellman operator and Gibbs map by

$$F_{h,s}(Q_h(s, \cdot)) := \frac{1}{\eta}\log\left\langle \pi_h^{\mathrm{ref}}(\cdot \mid s), e^{\eta Q_h(s,\cdot)}\right\rangle_{\mathcal{A}},$$

$$G_{h,s}(Q_h(s, \cdot))(a) := \frac{\pi_h^{\mathrm{ref}}(a \mid s)e^{\eta Q_h(s,a)}}{\sum_{a' \in \mathcal{A}} \pi_h^{\mathrm{ref}}(a' \mid s)e^{\eta Q_h(s,a')}}.$$

The Gibbs variational identity is

$$F_{h,s}(Q_h(s, \cdot)) = \max_{\pi_h(\cdot|s) \in \Delta(\mathcal{A})}\left\{\langle \pi_h(\cdot \mid s), Q_h(s, \cdot)\rangle_{\mathcal{A}} - \frac{1}{\eta}\mathrm{KL}\left(\pi_h(\cdot \mid s)\middle\|\pi_h^{\mathrm{ref}}(\cdot \mid s)\right)\right\},$$

with unique maximizer $G_{h,s}(Q_h(s, \cdot))$. For every h, s, define the optimal value function by $V_h^\star(s) := \sup_{\pi \in \Pi} V_h^\pi(s)$, where Π denotes the class of Markov policies, and let $V_{H+1}^\star \equiv 0$. These results follow from the standard characterization of KL-regularized Bellman optimality(Geist et al., 2019; Zhao et al., 2025b).

Proposition 2.1. *There exists a KL-regularized optimal policy $\pi^\star$, whose value and Q-functions satisfy, for every (h, s, a),*

$$Q_h^\star(s, a) = r_h(s, a) + (\mathbb{P}_h V_{h+1}^\star)(s, a), \qquad V_h^\star(s) = F_{h,s}(Q_h^\star(s, \cdot)),$$

with the optimal policy given by $\pi_h^\star(\cdot \mid s) = G_{h,s}(Q_h^\star(s, \cdot))$.

Moreover, $V_h^\pi(s) \le V_h^\star(s)$ and $Q_h^\pi(s, a) \le Q_h^\star(s, a)$ for every policy π, and the value ranges satisfy

$$0 \le V_h^\star(s) \le L_h, \qquad 0 \le Q_h^\star(s, a) \le L_h. \tag{1}$$

2.2 Offline Learning and Data Coverage

Offline Data. Let $\mu := \{\mu_h\}_{h=1}^H$ be a behavior policy. The offline dataset $D_\mu := \{\tau^k\}_{k=1}^K$ consists of K independent trajectories

$$\tau^k = (s_1^k, a_1^k, r_1^k, s_2^k, \ldots, s_H^k, a_H^k, r_H^k, s_{H+1}^k)$$

, where $s_1^k \sim \rho$, $a_h^k \sim \mu_h(\cdot \mid s_h^k)$, $s_{h+1}^k \sim \mathbb{P}_h(\cdot \mid s_h^k, a_h^k)$, and $r_h^k = r_h(s_h^k, a_h^k)$. The total number of transitions is $T = KH$. Since no further interaction with the environment is allowed, a finite-sample guarantee requires the offline data to provide sufficient coverage of the state-action pairs relevant to the target policy. We formalize this requirement through single-policy concentrability.

Single-policy Concentrability. Following prior analyses of offline RL under single-policy concentrability (Rashidinejad et al., 2021; Xie et al., 2021; Shi et al., 2022), we impose the following coverage assumption. It compares the occupancy measure of the KL-regularized optimal policy with that induced by the behavior policy. The coefficient $C_\eta^\star$ therefore measures how well the behavior policy covers the KL-regularized optimal policy $\pi^\star$, where the subscript η reflects the dependence of $\pi^\star$ and its occupancy measure on the regularization strength.

Assumption 2.2 (Single-policy concentrability)**.** *The behavior policy μ satisfies*

$$\max_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mu}(s,a)} \le C_\eta^\star$$

for some $C_\eta^\star \in [1,\infty)$, where $0/0 := 0$ and $x/0 := +\infty$ for $x > 0$.

Remark 2.1. Recall that μ is the behavior policy used to generate the offline dataset, while π^{ref} is the reference policy used in the KL regularization. We do not require μ to coincide with π^{ref}. Since the KL-regularized optimal policy $\pi^\star$ depends on η through the Gibbs optimality equations, its occupancy measure, and hence the concentrability coefficient $C_\eta^\star$, may also depend on η even when μ and π^{ref} are fixed.

Remark 2.2. In a finite unregularized MDP, there exists a deterministic optimal policy. Under KL regularization, however, $\mathrm{supp}(\pi_h^\star(\cdot \mid s)) = \mathrm{supp}(\pi_h^{\mathrm{ref}}(\cdot \mid s))$. Therefore, the single-policy concentrability condition may require coverage of more actions than in the unregularized case, where the optimal policy can be chosen to be deterministic.

Learning Objective. Our goal is to learn a policy π that maximizes the KL-regularized return $J_\eta(\pi) := \mathbb{E}_{s_1\sim\rho}[V_1^\pi(s_1)]$, and we measure the policy's performance by

$$\mathrm{SubOpt}(\pi) := J_\eta(\pi^\star) - J_\eta(\pi),$$

that is, the suboptimality of policy π under the initial distribution ρ. In addition, we say that a policy π is ϵ-optimal if $\mathrm{SubOpt}(\pi) \le \epsilon$.

3 Pessimism-Free KL-Regularized Q-Learning

We begin by introducing *KL-regularized Q-Learning (KLQ)*, a pessimism-free algorithm for the KL-regularized objective. Before proceeding, we first recall the classical Q-learning update (Watkins & Dayan, 1992). Classical Q-learning can be viewed as a stochastic approximation scheme for solving the Bellman optimality equation. Upon observing a transition (s_h, a_h, r_h, s_{h+1}) at step h, the corresponding Q-estimate is updated according to

$$Q_h(s_h,a_h) \leftarrow (1-\alpha_n)Q_h(s_h,a_h) + \alpha_n\big(r_h + V_{h+1}(s_{h+1})\big),$$

where α_n denotes the learning rate associated with the n-th visit to (s_h, a_h). In the standard unregularized setting, the value estimate is obtained by the hard maximum $V_h(s) = \max_{a\in\mathcal{A}} Q_h(s,a)$. For our KL-regularized objective, we retain the same stochastic approximation form of the Q-learning update and replace the hard maximum with the regularized Bellman operator introduced in Section 2. Specifically, we set $V_h(s) = F_{h,s}(Q_h(s,\cdot))$, and the policy associated with the Q estimate is given by $G_{h,s}(Q_h(s,\cdot))$. Like standard Q-learning, KLQ updates only the visited state-action pair as the offline trajectories are processed and does not construct an empirical transition model.

Algorithm 1 KL-Regularized Q-Learning (KLQ)

Require: Offline dataset D_μ, reference policy π^{ref}, inverse temperature $\eta > 0$

Initialize $Q_h^1(s,a) \leftarrow 0$ and $N_h^1(s,a) \leftarrow 0$ for every $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$

Set $V_h^1(s) \leftarrow F_{h,s}(Q_h^1(s,\cdot)) = 0$ for every $(h,s) \in [H] \times \mathcal{S}$, and set $V_{H+1}^k \equiv 0$

for $k = 1, \dots, K$ **do**

  Read the k-th trajectory τ^k from D_μ

  $Q^{k+1} \leftarrow Q^k$, $V^{k+1} \leftarrow V^k$, and $N^{k+1} \leftarrow N^k$

  for $h = 1, \dots, H$ **do**

    $N_h^{k+1}(s_h^k, a_h^k) \leftarrow N_h^k(s_h^k, a_h^k) + 1$

    $n \leftarrow N_h^{k+1}(s_h^k, a_h^k)$

    $\alpha_n \leftarrow \frac{H+1}{H+n}$

    $Q_h^{k+1}(s_h^k, a_h^k) \leftarrow (1-\alpha_n) Q_h^k(s_h^k, a_h^k) + \alpha_n (r_h^k + V_{h+1}^k(s_{h+1}^k))$

    $V_h^{k+1}(s_h^k) \leftarrow F_{h,s_h^k}(Q_h^{k+1}(s_h^k, \cdot))$

  end for

end for

Ensure: $\widehat{Q}_h \leftarrow Q_h^{K+1}$ and $\widehat{\pi}_h(\cdot \mid s) \leftarrow G_{h,s}(\widehat{Q}_h(s,\cdot))$

Algorithm 1 summarizes the resulting procedure. For the n-th visit to a state-action pair, we use the standard finite-horizon Q-learning step size $\alpha_n = \frac{H+1}{H+n}$ from Jin et al. (2018). After processing all K trajectories, the output policy is given by $\widehat{\pi}_h(\cdot \mid s) = G_{h,s}\big(\widehat{Q}_h(s,\cdot)\big)$ with $\widehat{Q}_h = Q_h^{K+1}$.

Compared with the pessimistic Q-learning algorithm of Shi et al. (2022), KLQ does not use an explicit pessimistic bonus or a monotone max update. While the algorithm of Shi et al. (2022) adjusts the Q-learning target by a pessimistic penalty and updates the value estimate monotonically, the Q update of KLQ is the same as in classical Q-learning, with the value function and policy determined by the KL-regularized Bellman equations. The following theorem gives a finite-sample guarantee for KLQ under single-policy concentrability.

Theorem 3.1. *Suppose Assumption 2.2 holds. For any $\delta \in (0,1)$, define $\iota := \log \frac{SAT}{\delta}$. Then, with probability at least $1-\delta$, the policy $\widehat{\pi}$ returned by Algorithm 1 satisfies*

$$\mathrm{SubOpt}(\widehat{\pi}) \le c \frac{\eta e^{\eta H(H+1)} H^7 S A C_\eta^\star}{K} \iota^4 (1 + \log K),$$

where $c > 0$ is a sufficiently large universal constant. As a consequence, KLQ is guaranteed to find an ϵ-optimal policy with sample complexity $\widetilde{O}\big(\eta e^{\eta H(H+1)} H^8 S A C_\eta^\star / \epsilon\big)$.

The key mechanism behind this fast dependence is that KL regularization makes policy suboptimality quadratic in the Q-estimation error, as formalized in Lemma C.2. However, the bound contains the exponential factor $e^{\eta H(H+1)}$, which can become large as the horizon increases. To control this factor, one may choose $\eta = \Theta(H^{-2})$, in which case $\eta H(H+1) = O(1)$ and the resulting sample complexity becomes $\widetilde{O}\big(H^6 S A C_\eta^\star / \epsilon\big)$. Such a choice, however, makes the inverse temperature depend on the horizon. Thus, an appropriate scaling of η with the horizon is required to obtain an explicit polynomial dependence on H, apart from the concentrability coefficient $C_\eta^\star$.

Remark 3.1. Similar exponential dependence appears in recent analyses of KL-regularized offline learning without pessimism. In single-step problems, the bound can depend exponentially on the inverse regularization strength (Wu et al., 2025; Chen & Zhang, 2026a), while in multi-step Markov games the corresponding factors may accumulate over the horizon (Chen et al., 2026). The same phenomenon arises in our analysis because we use a similar likelihood ratio argument based on the Gibbs form of KL-regularized policies. Specifically, this argument compares $G_{h,s}(Q_h^k(s,\cdot))$ with $\pi_h^\star(\cdot \mid s)$, producing exponential factors that accumulate over the horizon.

4 KL-Regularized Pessimistic Q-Learning

The exponential factor in Theorem 3.1 motivates us to consider a pessimistic variant of KLQ to improve the dependence on the regularization parameter and the horizon. We therefore introduce

Algorithm 2 KL-Regularized Pessimistic Q-Learning (KLPQ)

Require: Offline dataset D_μ, reference policy π^{ref}, inverse temperature $\eta > 0$, failure probability $\delta \in (0,1)$

Set $\iota \leftarrow \log \frac{SAT}{\delta}$ and $b_{h,n} \leftarrow c_b L_{h+1} \iota \sqrt{\frac{H}{n}}$ for a fixed sufficiently large universal constant c_b
Initialize $\widetilde{Q}_h^1(s,a) \leftarrow 0$, $Q_h^1(s,a) \leftarrow 0$, and $N_h^1(s,a) \leftarrow 0$ for every $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$
Set $V_h^1(s) \leftarrow F_{h,s}(Q_h^1(s,\cdot)) = 0$ for every $(h,s) \in [H] \times \mathcal{S}$, and set $V_{H+1}^k \equiv 0$
for $k = 1, \dots, K$ **do**
    Read the k-th trajectory τ^k from D_μ
    $\widetilde{Q}^{k+1} \leftarrow \widetilde{Q}^k$, $Q^{k+1} \leftarrow Q^k$, $V^{k+1} \leftarrow V^k$, and $N^{k+1} \leftarrow N^k$
    for $h = 1, \dots, H$ **do**
        $N_h^{k+1}(s_h^k, a_h^k) \leftarrow N_h^k(s_h^k, a_h^k) + 1$
        $n \leftarrow N_h^{k+1}(s_h^k, a_h^k)$
        $\alpha_n \leftarrow \frac{H+1}{H+n}$
        $\widetilde{Q}_h^{k+1}(s_h^k, a_h^k) \leftarrow (1-\alpha_n)\widetilde{Q}_h^k(s_h^k, a_h^k) + \alpha_n(r_h^k + V_{h+1}^k(s_{h+1}^k) - b_{h,n})$
        $Q_h^{k+1}(s_h^k, a_h^k) \leftarrow \max\{Q_h^k(s_h^k, a_h^k), \widetilde{Q}_h^{k+1}(s_h^k, a_h^k)\}$
        $V_h^{k+1}(s_h^k) \leftarrow F_{h,s_h^k}(Q_h^{k+1}(s_h^k, \cdot))$
    end for
end for

Ensure: $\widehat{Q}_h \leftarrow Q_h^{K+1}$ and $\widehat{\pi}_h(\cdot \mid s) \leftarrow G_{h,s}(\widehat{Q}_h(s,\cdot))$

KL-regularized Pessimistic Q-Learning (KLPQ). KLPQ retains the regularized value and policy updates of KLQ, while modifying the Q-estimation procedure through a pessimistic bonus and a monotone update. More specifically, KLPQ introduces the pessimistic bonus

$$b_{h,n} := c_b L_{h+1} \iota \sqrt{\frac{H}{n}}, \qquad \iota := \log \frac{SAT}{\delta}, \qquad n \geq 1, \tag{2}$$

where $c_b > 0$ is a fixed sufficiently large universal constant. In KLPQ, we maintain an auxiliary Q estimate $\widetilde{Q}^k$ in addition to Q^k. At the n-th visit to (s_h^k, a_h^k), we first update $\widetilde{Q}_h^{k+1}(s_h^k, a_h^k)$ using the pessimistic Q-learning target:

$$\widetilde{Q}_h^{k+1}(s_h^k, a_h^k) \leftarrow (1-\alpha_n)\widetilde{Q}_h^k(s_h^k, a_h^k) + \alpha_n\big(r_h^k + V_{h+1}^k(s_{h+1}^k) - b_{h,n}\big).$$

Inspired by Shi et al. (2022), we use the monotone max update

$$Q_h^{k+1}(s_h^k, a_h^k) \leftarrow \max\left\{Q_h^k(s_h^k, a_h^k), \widetilde{Q}_h^{k+1}(s_h^k, a_h^k)\right\},$$

which takes the maximum of the previous Q estimate and the updated auxiliary estimate. The two Q estimates serve different roles. The auxiliary iterate $\widetilde{Q}^k$ follows the Q-learning update with the pessimistic bonus, while Q^k records the monotone estimate used to compute the regularized value function. In the unregularized pessimistic Q-learning algorithm of Shi et al. (2022), monotonicity is enforced directly at the value level by taking the maximum of the previous value estimate and $\max_a Q_h(s,a)$. In our regularized setting, the value estimate is instead given by $F_{h,s}(Q_h(s,\cdot))$. Accordingly, we apply the max update to the Q estimate and then compute the value through $F_{h,s}$. Since $F_{h,s}$ is monotone in each coordinate, this also makes the resulting value estimates monotone, which is crucial for our theoretical analysis. For each $k \in [K+1]$, let π^k denote the policy induced by Q^k, i.e., $\pi_h^k(\cdot \mid s) := G_{h,s}(Q_h^k(s,\cdot))$. In particular, $\widehat{\pi} = \pi^{K+1}$. Algorithm 2 summarizes the resulting procedure. Under the same single-policy concentrability, KLPQ admits the following finite-sample guarantee.

Theorem 4.1. *Suppose Assumption 2.2 holds. For any $\delta \in (0,1)$, the policy $\widehat{\pi}$ returned by Algorithm 2 satisfies, with probability at least $1-\delta$,*

$$\mathrm{SubOpt}(\widehat{\pi}) \leq c\frac{\eta H^6 S A C_\eta^\star}{K}\iota^3(1+\log K),$$

where $c > 0$ is a universal constant. As a consequence, KLPQ is guaranteed to find an ϵ-optimal policy with sample complexity $\widetilde{O}\big(\eta H^7 S A C_\eta^\star/\epsilon\big)$.

Consequently, η can be fixed as H increases without introducing an explicit exponential dependence on the horizon, yielding a sample complexity of $\widetilde{O}\big(H^7SAC^\star_\eta/\epsilon\big)$ for fixed η. In comparison, under the same scaling $\eta = \Theta(H^{-2})$ considered for KLQ, the sample complexity becomes $\widetilde{O}\big(H^5SAC^\star_\eta/\epsilon\big)$, improving the corresponding H^6 dependence of KLQ by one factor of H. Apart from the concentrability coefficient $C^\star_\eta$, KLPQ therefore retains an explicit polynomial dependence on H without requiring η to scale with H. Both improvements rely on the pessimistic and monotone estimates maintained by KLPQ, we give the analysis in Section 5.

Remark 4.1. A similar role of pessimism has also been observed in KL-regularized offline bandits, where Ji et al. (2026a) construct a pessimistic reward estimate and obtain polynomial rather than exponential dependence on η. Their setting is single-step, so pessimism is imposed directly on the reward estimate before forming the Gibbs policy. KLPQ, by contrast, maintains pessimism throughout the multi-step Q-learning updates. Section 5 gives the corresponding sequential analysis.

Remark 4.2. Explicit exponential dependence can also be avoided without pessimism when additional structure is available. For example, in single-step zero-sum games, Zhang et al. (2026) exploit equilibrium stability and the skew-symmetry of the game operator to localize the analysis to unilateral estimation errors. Their argument relies on a problem specific setting. In the multi-step Q-learning problem considered here, we instead use pessimism to remove the exponential factor.

Comparison with Q-learning under the Unregularized Setting. Shi et al. (2022) study offline Q-learning in finite-horizon tabular MDPs under single-policy concentrability, but for the standard unregularized objective. They achieve a sample complexity of $\widetilde{O}(H^4SC^\star/\epsilon^2 + H^5SC^\star/\epsilon)$, with its leading ϵ^{-2} term matching the minimax lower bound $\widetilde{\Omega}(H^4SC^\star/\epsilon^2)$ up to logarithmic factors (Xie et al., 2021). Our guarantees concern a different, KL-regularized objective. Under KL regularization, the optimal policy is generally stochastic, whereas an optimal policy under the standard objective can be chosen to be deterministic. Consequently, single-policy concentrability may require coverage of more actions, which is reflected in the additional explicit factor A in our bounds. More importantly, KL regularization changes the statistical geometry of the learning problem and permits policy suboptimality to depend quadratically on Q-estimation error. Accordingly, both KLQ and KLPQ achieve an ϵ^{-1} dependence. In particular, KLPQ achieves $\widetilde{O}(\eta H^7SAC^\star_\eta/\epsilon)$. This improvement in the dependence on ϵ does not contradict the minimax lower bound for the unregularized objective, but instead reflects the statistical benefit introduced by the KL regularization.

5 Theoretical Analysis

In this section, we outline the main steps needed to establish Theorems 3.1 and 4.1. We begin with the analysis of KLQ, which also provides the basic ingredients for the analysis of KLPQ. We then turn to KLPQ and highlight the changes brought by the pessimistic construction. The detailed proofs are deferred to Appendices C and D.

5.1 Analysis for KLQ

We begin by introducing some notation. Recall that $L_h := H - h + 1$. For $1 \le n \le N$, let $\alpha_n^N := \alpha_n \prod_{j=n+1}^N (1-\alpha_j)$. For each iterate Q_h^k, we define $e_h^k(s) := \mathbb{E}_{a\sim\pi_h^\star(\cdot|s)}[(Q_h^k(s,a) - Q_h^\star(s,a))^2]$ and set $e_{H+1}^k \equiv 0$. We also write $\gamma_h := e^{2\eta L_h}$, with $\gamma_{H+1} := 1$, and $\Gamma_{a:b} := \prod_{j=a}^b \gamma_j$, where $\Gamma_{a:b} := 1$ when $a > b$. In particular, $\Gamma_{1:H} = e^{\eta H(H+1)}$.

Proof Sketch of Theorem 3.1. Lemma C.2 first relates the policy suboptimality to the squared Q-estimation error: $\mathrm{SubOpt}(\widehat{\pi}) \le \frac{\eta}{2}\sum_{h=1}^H \Gamma_{1:h}\mathbb{E}_{(s,a)\sim d_h^{\pi^\star}}[(Q_h^{K+1}(s,a) - Q_h^\star(s,a))^2]$. It therefore remains to control the squared error of the final Q iterate.

Step 1: Decomposing the Q-estimation Error. Fix (s,a) at step h, let $N := N_h^k(s,a) \ge 1$, and let $k_{h,n} = k_{h,n}(s,a)$ denote the episode in which (s,a) is visited for the n-th time at step h. Unrolling the Q-learning update and subtracting the optimal Bellman equation gives $Q_h^k(s,a) - Q_h^\star(s,a) = \mathrm{Term(i)} + \mathrm{Term(ii)}$, where $\mathrm{Term(i)} := \sum_{n=1}^N \alpha_n^N[V_{h+1}^{k_{h,n}}(s_{h+1}^{k_{h,n}}) - (\mathbb{P}_h V_{h+1}^{k_{h,n}})(s,a)]$ and $\mathrm{Term(ii)} := \sum_{n=1}^N \alpha_n^N(\mathbb{P}_h(V_{h+1}^{k_{h,n}} - V_{h+1}^\star))(s,a)$. Here, Term(i) captures the transition sampling error, while Term(ii) propagates the value-estimation error from stage

$h+1$. By Lemma B.3, we have $|\text{Term(i)}| \le cL_{h+1}\iota\sqrt{H/N}$. For Term(ii), Lemma C.2 gives $(V^k_{h+1}(s) - V^\star_{h+1}(s))^2 \le \gamma_{h+1}e^k_{h+1}(s)$. Squaring the decomposition and applying Jensen's inequality to Term(ii), as detailed in the proof of Lemma C.3, gives $(Q^k_h(s,a) - Q^\star_h(s,a))^2 \le c_0H^2L^2_{h+1}\iota^2/N + (1+1/H)\gamma_{h+1}\sum_{n=1}^N \alpha^N_n(\mathbb{P}_h e^{k_{h,n}}_{h+1})(s,a)$. Thus, it remains to control the propagated next-stage error after summing this pointwise bound over episodes.

Step 2: Establishing the Key Recursion. We sum the bound in Lemma C.3 over $k \in [K]$ and take expectation over $(s,a) \sim d^{\pi^\star}_h$. The remaining difficulty is to control the historical next-stage error term. Lemma C.4 shows that, after summing over episodes, this term can be related to the cumulative squared error at stage $h+1$, yielding $\sum_{k=1}^K \sum_s d^{\pi^\star}_h(s)e^k_h(s) \le (1+5/H)\gamma_{h+1}\sum_{k=1}^K\sum_s d^{\pi^\star}_{h+1}(s)e^k_{h+1}(s) + c_0SAC^\star_\eta\gamma_{h+1}H^2L^2_h\iota^3(1+\log K)$. Thus, we obtain a recursion relating the cumulative squared errors at stages h and $h+1$. Iterating this recursion backward from H to h gives $\sum_{k=1}^K\sum_s d^{\pi^\star}_h(s)e^k_h(s) \le c_1SAC^\star_\eta\Gamma_{h+1:H}H^2L^3_h\iota^3(1+\log K)$.

Step 3: Controlling the Final Iterate. Combining Lemmas C.5 and C.4 gives, up to a universal constant, $\mathbb{E}_{(s,a)\sim d^{\pi^\star}_h}\left[(Q^{K+1}_h(s,a) - Q^\star_h(s,a))^2\right] \le \widetilde{O}\left(\frac{SAC^\star_\eta\Gamma_{h+1:H}H^3L^3_h}{K}\right)$. Together with $\text{SubOpt}(\widehat{\pi}) \le \frac{\eta}{2}\sum_{h=1}^H \Gamma_{1:h}\mathbb{E}_{d^{\pi^\star}_h}[(Q^{K+1}_h - Q^\star_h)^2]$ and $\Gamma_{1:h}\Gamma_{h+1:H} = \Gamma_{1:H} = e^{\eta H(H+1)}$, this proves Theorem 3.1 after collecting the polynomial and logarithmic factors. The product $\Gamma_{h+1:H}$ in Lemma C.4 comes from the factors γ_{h+1} accumulated through the stagewise recursion, together with $\Gamma_{1:h}$ in Lemma C.2, it gives the exponential dependence on η and H.

5.2 Analysis for KLPQ

Proof Sketch of Theorem 4.1. The proof follows the same overall route as the KLQ analysis: a pointwise Q-estimation error bound is first converted into a stagewise recursion and then into a bound for the final iterate. The key difference lies in how to control the propagated next stage error. Under the pessimistic construction, the factor γ_{h+1} appearing in the KLQ recursion can be removed without changing the order of the statistical error. To see this, the role of the pessimistic bonus in the analysis is twofold. First, together with the monotone update, it yields the pessimistic ordering $Q^k_h(s,a) \le Q^{\pi^k}_h(s,a) \le Q^\star_h(s,a)$ by Lemma D.2. By this one-sided ordering, Lemma D.3 yields, with high probability, $(V^\star_h(s) - V^k_h(s))^2 \le e^k_h(s)$ and $\text{SubOpt}(\widehat{\pi}) \le \eta\sum_{h=1}^H \mathbb{E}_{d^{\pi^\star}_h}[(Q^\star_h - Q^{K+1}_h)^2]$, without the factor γ_h required in the corresponding KLQ bounds. Second, the pessimistic bonus is calibrated so that its accumulated contribution satisfies $\sum_{n=1}^N \alpha^N_n b_{h,n} \asymp L_{h+1}\iota\sqrt{H/N}$, matching the order of the transition sampling error in Lemma B.3. Hence introducing pessimism does not change the order of the statistical error in the pointwise recursion. Using $(V^\star_{h+1} - V^k_{h+1})^2 \le e^k_{h+1}$, Lemma D.4 and the same summation argument give $\sum_{k=1}^K\sum_s d^{\pi^\star}_h(s)e^k_h(s) \le (1+5/H)\sum_{k=1}^K\sum_s d^{\pi^\star}_{h+1}(s)e^k_{h+1}(s) + cSAC^\star_\eta H^2L^2_h\iota^3(1+\log K)$. Iterating this recursion over h therefore introduces no product of the factors γ_h. Finally, monotonicity further simplifies the control of the final iterate. Since the Q-estimates increase monotonically while remaining below $Q^\star$, the final iterate satisfies $0 \le Q^\star_h - Q^{K+1}_h \le Q^\star_h - Q^k_h$ for every k. Therefore, $\mathbb{E}_{d^{\pi^\star}_h}[(Q^{K+1}_h - Q^\star_h)^2] \le K^{-1}\sum_{k=1}^K \mathbb{E}_{d^{\pi^\star}_h}[(Q^k_h - Q^\star_h)^2]$. Combining this inequality with the stagewise bound above and Lemma D.3 proves Theorem 4.1, while avoiding the additional final iterate analysis required for KLQ. The detailed proof is deferred to Appendix D.

6 Discussion

Recent work on offline KL-regularized multi-armed bandits has established matching upper and lower bounds across different regularization regimes, showing that the minimax optimal sample complexity itself depends on the strength of regularization (Ji et al., 2026a). In particular, the optimal rate exhibits a phase transition between a fast rate regime under sufficiently strong regularization and the slower rate of the unregularized problem as the regularization weakens. Whether a similar transition occurs in finite-horizon MDPs remains open. Establishing matching lower bounds across different regularization regimes would therefore be important for determining when the fast rates established in this work are statistically optimal.

AI USE STATEMENT

We used generative AI tools to assist with language polishing, literature retrieval and discovery, drafting portions of the manuscript, and developing and checking the mathematical proofs. All AI-assisted literature suggestions were independently verified against the original sources, and all AI-assisted mathematical arguments and manuscript text were reviewed, revised, and validated by the authors. The authors take full responsibility for the final content of the paper, including its mathematical claims, proofs, citations, and presentation.

REFERENCES

Claire Chen and Yuheng Zhang. Fast rates in α-potential games via regularized mirror descent. *arXiv preprint arXiv:2605.00268*, 2026a.

Claire Chen and Yuheng Zhang. Pessimism-free offline learning in general-sum games via kl regularization. *arXiv preprint arXiv:2605.00264*, 2026b.

Claire Chen, Yuheng Zhang, Xinyu Liu, Zixuan Xie, Shuze Daniel Liu, and Nan Jiang. Offline two-player zero-sum markov games with KL regularization. In *Forty-third International Conference on Machine Learning*, 2026.

Eyal Even-Dar and Yishay Mansour. Learning rates for q-learning. *Journal of machine learning Research*, 5(Dec):1–25, 2003.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118, 1975.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pp. 2052–2062. PMLR, 2019.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International conference on machine learning*, pp. 2160–2169. PMLR, 2019.

Haoyang Hong, Zichen Wang, Quanquan Gu, and Huazheng Wang. Online kl-regularized reinforcement learning with function approximation under misspecification. *arXiv preprint arXiv:2606.06053*, 2026.

Kaixuan Ji, Qiwei Di, Heyang Zhao, Qingyue Zhao, and Quanquan Gu. On the optimal sample complexity of offline multi-armed bandits with kl regularization. *arXiv preprint arXiv:2605.02141*, 2026a.

Kaixuan Ji, Qingyue Zhao, Heyang Zhao, Qiwei Di, and Quanquan Gu. Near-optimal regret for kl-regularized multi-armed bandits. *arXiv preprint arXiv:2603.02155*, 2026b.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International conference on machine learning*, pp. 5084–5096. PMLR, 2021.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33:1179–1191, 2020.

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

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous q-learning: Sharper analysis and variance reduction. *Advances in neural information processing systems*, 33:7031–7043, 2020.

Gen Li, Changxiao Cai, Yuxin Chen, Yuantao Gu, Yuting Wei, and Yuejie Chi. Tightening the dependence on horizon in the sample complexity of q-learning. In *International Conference on Machine Learning*, pp. 6296–6306. PMLR, 2021a.

Gen Li, Laixi Shi, Yuxin Chen, Yuantao Gu, and Yuejie Chi. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Advances in Neural Information Processing Systems*, 34:17762–17776, 2021b.

Gen Li, Changxiao Cai, Yuxin Chen, Yuting Wei, and Yuejie Chi. Is q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*, 72(1):222–236, 2024.

Xinyu Liu, Zixuan Xie, and Shangtong Zhang. Linear Q-learning does not diverge in L^2: Convergence rates to a bounded set. In *Forty-second International Conference on Machine Learning*, 2025.

Anupam Nayak, Tong Yang, Osman Yagan, Gauri Joshi, and Yuejie Chi. Achieving logarithmic regret in kl-regularized zero-sum markov games. *arXiv preprint arXiv:2510.13060*, 2025.

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

Guannan Qu and Adam Wierman. Finite-time analysis of asynchronous stochastic approximation and q-learning. In *Conference on learning theory*, pp. 3185–3205. PMLR, 2020.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International conference on machine learning*, pp. 19967–20025. PMLR, 2022.

Daniil Tiapkin, Denis Belomestny, Daniele Calandriello, Eric Moulines, Remi Munos, Alexey Naumov, Pierre Perrault, Yunhao Tang, Michal Valko, and Pierre Menard. Fast rates for maximum entropy exploration. In *International Conference on Machine Learning*, pp. 34161–34221. PMLR, 2023.

Daniil Tiapkin, Denis Belomestny, Daniele Calandriello, Eric Moulines, Alexey Naumov, Pierre Perrault, Michal Valko, and Pierre Menard. Demonstration-regularized RL. In *The Twelfth International Conference on Learning Representations*, 2024.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992.

Di Wu, Chengshuai Shi, Jing Yang, and Cong Shen. Greedy sampling is provably efficient for RLHF. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

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

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. *arXiv preprint arXiv:2312.11456*, 2023.

Yuling Yan, Gen Li, Yuxin Chen, and Jianqing Fan. The efficacy of pessimism in asynchronous q-learning. *IEEE Transactions on Information Theory*, 69(11):7185–7219, 2023.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pp. 2730–2775. PMLR, 2022.

Yuheng Zhang, Claire Chen, and Nan Jiang. Beyond pessimism: Offline learning in kl-regularized games. *arXiv preprint arXiv:2604.06738*, 2026.

Heyang Zhao, Chenlu Ye, Quanquan Gu, and Tong Zhang. Sharp analysis for KL-regularized contextual bandits and RLHF. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Heyang Zhao, Chenlu Ye, Wei Xiong, Quanquan Gu, and Tong Zhang. Logarithmic regret for online kl-regularized reinforcement learning. *arXiv preprint arXiv:2502.07460*, 2025b.

Qingyue Zhao, Kaixuan Ji, Heyang Zhao, and Quanquan Gu. Fast rates for offline contextual bandits with forward-kl regularization under single-policy concentrability. *arXiv preprint arXiv:2605.09214*, 2026a.

Qingyue Zhao, Kaixuan Ji, Heyang Zhao, Tong Zhang, and Quanquan Gu. Towards a sharp analysis of offline policy learning for f-divergence-regularized contextual bandits. In *The Fourteenth International Conference on Learning Representations*, 2026b.

APPENDIX

A Proof of Proposition 2.1 **14**

B Supporting Lemmas **14**

C Proofs for Algorithm 1 **17**

C.1 Lemmas for Theorem 3.1 17

C.2 Proof of Theorem 3.1 23

D Proofs for Algorithm 2 **24**

D.1 Lemmas for Theorem 4.1 24

D.2 Proof of Theorem 4.1 28

A Proof of Proposition 2.1

Proof. Define $\bar{V}_{H+1} \equiv 0$ and recursively set $\bar{Q}_h := r_h + \mathbb{P}_h \bar{V}_{h+1}$, $\bar{V}_h(s) := F_{h,s}(\bar{Q}_h(s,\cdot))$, and $\bar{\pi}_h(\cdot \mid s) := G_{h,s}(\bar{Q}_h(s,\cdot))$. We prove by backward induction that $V_h^\pi \le \bar{V}_h$ for every policy π and $V_h^{\bar{\pi}} = \bar{V}_h$. Both claims are immediate at $H+1$. If $V_{h+1}^\pi \le \bar{V}_{h+1}$, then

$$\begin{aligned} V_h^\pi(s) &= \big\langle \pi_h(\cdot \mid s), r_h(s,\cdot) + (\mathbb{P}_h V_{h+1}^\pi)(s,\cdot) \big\rangle_{\mathcal{A}} - \frac{1}{\eta}\mathrm{KL}\big(\pi_h(\cdot \mid s) \big\| \pi_h^{\mathrm{ref}}(\cdot \mid s)\big) \\ &\le \big\langle \pi_h(\cdot \mid s), \bar{Q}_h(s,\cdot) \big\rangle_{\mathcal{A}} - \frac{1}{\eta}\mathrm{KL}\big(\pi_h(\cdot \mid s) \big\| \pi_h^{\mathrm{ref}}(\cdot \mid s)\big) \\ &\le F_{h,s}(\bar{Q}_h(s,\cdot)) = \bar{V}_h(s), \end{aligned}$$

where the last inequality is the Gibbs variational identity. If $V_{h+1}^{\bar{\pi}} = \bar{V}_{h+1}$, then $Q_h^{\bar{\pi}} = \bar{Q}_h$, and because $\bar{\pi}_h$ is the maximizer in the Gibbs variational identity,

$$V_h^{\bar{\pi}}(s) = \big\langle \bar{\pi}_h(\cdot \mid s), \bar{Q}_h(s,\cdot) \big\rangle_{\mathcal{A}} - \frac{1}{\eta}\mathrm{KL}\big(\bar{\pi}_h(\cdot \mid s) \big\| \pi_h^{\mathrm{ref}}(\cdot \mid s)\big) = \bar{V}_h(s).$$

Thus $\bar{V}_h = V_h^\star$, $\bar{Q}_h = Q_h^\star$, and $\bar{\pi} = \pi^\star$. The same induction also gives $Q_h^\pi \le Q_h^\star$ for every policy π. For the range, suppose $0 \le V_{h+1}^\star \le L_{h+1}$. Then

$$0 \le Q_h^\star(s,a) = r_h(s,a) + (\mathbb{P}_h V_{h+1}^\star)(s,a) \le 1 + L_{h+1} = L_h.$$

Since $F_{h,s}(q)$ lies between the minimum and maximum coordinates of q,

$$0 \le V_h^\star(s) = F_{h,s}(Q_h^\star(s,\cdot)) \le L_h.$$

Backward induction from $V_{H+1}^\star = 0$ completes the proof. □

B Supporting Lemmas

We first introduce the notation used throughout the statistical analysis. Let $N_h^k(s,a)$ denote the number of visits to (h,s,a) before episode k, and let $k_{h,n}(s,a)$ denote the episode index of its n-th visit. For the learning rate $\alpha_n = \frac{H+1}{H+n}$ used in Algorithm 1, for integers $N \ge 0$ and $n \ge 1$, define

$$\alpha_0^N := \begin{cases} \prod_{j=1}^{N}(1-\alpha_j) = 0, & N > 0, \\ 1, & N = 0, \end{cases} \qquad \alpha_n^N := \begin{cases} \alpha_n \prod_{j=n+1}^{N}(1-\alpha_j), & N > n, \\ \alpha_n, & N = n, \\ 0, & N < n. \end{cases}$$

These coefficients and their basic identities are standard in finite-sample analyses of Q-learning, see Jin et al. (2018, Lemma 4.1) and Li et al. (2021b, Lemma 1).

Let $\mathcal{D}_{<k} := \sigma(\tau^1, \dots, \tau^{k-1})$. Let $\widehat{\mathbb{P}}_h^k$ be the one-hot row vector satisfying $\widehat{\mathbb{P}}_h^k f = f(s_{h+1}^k)$, and regard $\widehat{\mathbb{P}}_h^k$ and $\mathbb{P}_{h,s,a}$ as row vectors over $\mathcal{S}$. Under Assumption 2.2,

$$0 \le \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mu}(s,a)} \le C_\eta^\star,$$

with the convention $0/0 = 0$, as in Assumption 2.2. Throughout the appendix, we use the convention that empty sums are zero. We record the standard results below.

Lemma B.1. *For every $N \ge 1$ and $a \in [1/2, 1]$,*

$$\frac{1}{N^a} \le \sum_{n=1}^{N} \frac{\alpha_n^N}{n^a} \le \frac{2}{N^a}, \qquad \sum_{n=1}^{N} \alpha_n^N = 1, \qquad \max_{1 \le n \le N} \alpha_n^N \le \frac{2H}{N},$$

$$\sum_{n=1}^{N} (\alpha_n^N)^2 \le \frac{2H}{N}, \qquad \sum_{M=n}^{\infty} \alpha_n^M = 1 + \frac{1}{H}.$$

We also need the following standard consequence of Freedman's martingale inequality (Freedman, 1975, Theorem 1.6).

Lemma B.2. *Let $X_1, \dots, X_m$ be martingale differences with respect to a filtration $\{\mathcal{F}_i\}_{i=0}^m$, let $|X_i| \le R$, and set $W_m := \sum_{i=1}^m \mathbb{E}[X_i^2 \mid \mathcal{F}_{i-1}]$. If $W_m \le \sigma^2$ almost surely for a deterministic σ^2, then, for every $u > 0$, with probability at least $1 - 2e^{-u}$,*

$$\left|\sum_{i=1}^m X_i\right| \le c\sqrt{\sigma^2 u} + cRu.$$

Lemma B.3. *Suppose that, for every $k \in [K]$ and $h \in [H]$, V_{h+1}^k is $\mathcal{D}_{<k}$-measurable and satisfies $0 \le V_{h+1}^k(s) \le L_{h+1}$ for every $s \in \mathcal{S}$. For any $\delta \in (0,1)$, let $\iota := \log \frac{SAT}{\delta}$. With probability at least $1-\delta$, simultaneously for every $h \in [H]$, $(s,a) \in \mathcal{S} \times \mathcal{A}$, and $1 \le N \le N_h^{K+1}(s,a)$,*

$$\left|\sum_{n=1}^N \alpha_n^N (\widehat{\mathbb{P}}_h^{k_{h,n}(s,a)} - \mathbb{P}_{h,s,a}) V_{h+1}^{k_{h,n}(s,a)}\right| \le cL_{h+1}\iota\sqrt{\frac{H}{N}},$$

where $c > 0$ is a universal constant.

Proof. If $L_{h+1} = 0$, the claim is immediate. Hence suppose $L_{h+1} > 0$. Then $H \ge 2$, so $SAT \ge 2$ and $\iota \ge \log 2$. Fix (h, s, a, N) with $1 \le N \le K$ and define

$$\overline{N}_h^k(s,a) := N_h^k(s,a) + \mathbf{1}\{(s_h^k, a_h^k) = (s,a)\}, \quad \mathcal{F}_{k,h}^- := \sigma(\mathcal{D}_{<k}, s_1^k, a_1^k, \dots, s_h^k, a_h^k),$$
$$\mathcal{F}_{k,h}^+ := \mathcal{F}_{k,h}^- \vee \sigma(s_{h+1}^k), \quad X_k := \alpha^N_{\overline{N}_h^k(s,a)} \mathbf{1}\{(s_h^k, a_h^k) = (s,a)\}(\widehat{\mathbb{P}}_h^k - \mathbb{P}_{h,s,a}) V_{h+1}^k.$$

Conditional on $\mathcal{F}_{k,h}^-$, every factor in X_k except s_{h+1}^k is fixed, and $s_{h+1}^k \sim \mathbb{P}_h(\cdot \mid s_h^k, a_h^k)$. Hence $\mathbb{E}[X_k \mid \mathcal{F}_{k,h}^-] = 0$. Moreover, $\mathcal{F}_{k,h}^- \subseteq \mathcal{F}_{k,h}^+ \subseteq \mathcal{F}_{k+1,h}^-$ for $k < K$. Define $\mathcal{G}_0 := \mathcal{F}_{1,h}^-$, $\mathcal{G}_k := \mathcal{F}_{k+1,h}^-$ for $1 \le k < K$, and $\mathcal{G}_K := \mathcal{F}_{K,h}^+$. Then $\{\mathcal{G}_k\}_{k=0}^K$ is a filtration, X_k is $\mathcal{G}_k$-measurable, and $\mathbb{E}[X_k \mid \mathcal{G}_{k-1}] = \mathbb{E}[X_k \mid \mathcal{F}_{k,h}^-] = 0$ for every $k \in [K]$. At the n-th visit to (s,a), $\overline{N}_h^k(s,a) = n$. Since $\alpha_n^N = 0$ for $n > N$, using $0 \le V_{h+1}^k \le L_{h+1}$ and Lemma B.1, we have $|X_k| \le \frac{2HL_{h+1}}{N}$. Furthermore,

$$\sum_{k=1}^K \mathbb{E}[X_k^2 \mid \mathcal{G}_{k-1}] = \sum_{k=1}^K \mathbb{E}[X_k^2 \mid \mathcal{F}_{k,h}^-] \le L_{h+1}^2 \sum_{k=1}^K \big(\alpha^N_{\overline{N}_h^k(s,a)}\big)^2 \mathbf{1}\{(s_h^k, a_h^k) = (s,a)\}$$
$$\le L_{h+1}^2 \sum_{n=1}^N (\alpha_n^N)^2 \le \frac{2HL_{h+1}^2}{N}.$$

Hence, for $N \ge H$, applying Lemma B.2 with $u = 2\iota$ yields, with probability at least $1 - 2e^{-2\iota}$,

$$\left|\sum_{k=1}^K X_k\right| \le cL_{h+1}\sqrt{\frac{H\iota}{N}} + cL_{h+1}\frac{H\iota}{N} \le cL_{h+1}\iota\sqrt{\frac{H}{N}},$$

where the last inequality uses $\iota \ge \log 2$ and $H/N \le \sqrt{H/N}$, after absorbing numerical factors into the universal constant. Whenever $N \le N_h^{K+1}(s,a)$, by the definition of the visit indices,

$$\sum_{k=1}^K X_k = \sum_{n=1}^N \alpha_n^N (\widehat{\mathbb{P}}_h^{k_{h,n}(s,a)} - \mathbb{P}_{h,s,a}) V_{h+1}^{k_{h,n}(s,a)}.$$

Thus the desired bound holds for $N \ge H$. When $N < H$ and $N \le N_h^{K+1}(s,a)$, the same quantity is deterministically bounded by

$$\left|\sum_{n=1}^N \alpha_n^N (\widehat{\mathbb{P}}_h^{k_{h,n}(s,a)} - \mathbb{P}_{h,s,a}) V_{h+1}^{k_{h,n}(s,a)}\right| \le L_{h+1}\sum_{n=1}^N \alpha_n^N = L_{h+1} \le cL_{h+1}\iota\sqrt{\frac{H}{N}},$$

where we used the weight normalization in Lemma B.1, $\iota \ge \log 2$, and $N < H$, after enlarging c if necessary. A union bound over at most $HSAK = SAT$ choices of (h, s, a, N) has failure probability at most $2SATe^{-2\iota} \le \delta$, proving the lemma. □

Lemma B.4. *Suppose $SAT > 1$. For any $\delta \in (0,1)$, let $\iota := \log \frac{SAT}{\delta}$. With probability at least $1-\delta$, simultaneously for every $h \in [H]$, $(s,a) \in \mathcal{S} \times \mathcal{A}$, and $m \in [K]$,*

$$\frac{d_h^\mu(s,a)}{N_h^{m+1}(s,a) \vee 1} \le \frac{8\iota}{m}, \tag{3}$$

and whenever $m d_h^\mu(s,a) \ge 8\iota$,

$$N_h^{m+1}(s,a) \ge \frac{m d_h^\mu(s,a)}{2} \ge 1. \tag{4}$$

Under Assumption 2.2, on the same event, simultaneously for every $h \in [H]$,

$$\sum_{k=1}^{K} \sum_{s,a} d_h^{\pi^\star}(s,a) \mathbf{1}\{N_h^k(s,a) = 0\} \le 1 + 8SAC_\eta^\star \iota, \tag{5}$$

$$\sum_{k=1}^{K} \sum_{s,a} d_h^{\pi^\star}(s,a) \frac{\mathbf{1}\{N_h^k(s,a) \ge 1\}}{N_h^k(s,a) \vee 1} \le 8SAC_\eta^\star \iota (1 + \log K), \tag{6}$$

$$\sum_{s,a} d_h^{\pi^\star}(s,a) \mathbf{1}\{N_h^{K+1}(s,a) = 0\} \le \frac{8SAC_\eta^\star \iota}{K}, \tag{7}$$

and

$$\sum_{s,a} d_h^{\pi^\star}(s,a) \frac{\mathbf{1}\{N_h^{K+1}(s,a) \ge 1\}}{N_h^{K+1}(s,a) \vee 1} \le \frac{8SAC_\eta^\star \iota}{K}. \tag{8}$$

Proof. Fix (h,s,a,m). The count $N_h^{m+1}(s,a)$ is binomial with parameters m and $d_h^\mu(s,a)$. If $m d_h^\mu(s,a) < 8\iota$, then

$$\frac{d_h^\mu(s,a)}{N_h^{m+1}(s,a) \vee 1} \le d_h^\mu(s,a) < \frac{8\iota}{m}.$$

Otherwise, the multiplicative Chernoff bound gives

$$\Pr\left(N_h^{m+1}(s,a) < \frac{m d_h^\mu(s,a)}{2} \right) \le e^{-m d_h^\mu(s,a)/8} \le e^{-\iota}.$$

On the complementary event, $N_h^{m+1}(s,a) \ge m d_h^\mu(s,a)/2 \ge 1$, and hence

$$\frac{d_h^\mu(s,a)}{N_h^{m+1}(s,a) \vee 1} \le \frac{2}{m} \le \frac{8\iota}{m}.$$

A union bound over at most $HSAK = SAT$ choices proves the first two claims of Lemma B.4 simultaneously.

To prove (5), observe that (4) with $m = k-1$ implies, for $k \ge 2$, $N_h^k(s,a) = 0$ implies $(k-1) d_h^\mu(s,a) < 8\iota$. Thus a state-action pair with positive behavior occupancy can have zero count for at most $1 + 8\iota / d_h^\mu(s,a)$ episodes. Using Assumption 2.2,

$$\sum_{k=1}^{K} \sum_{s,a} d_h^{\pi^\star}(s,a) \mathbf{1}\{N_h^k(s,a) = 0\} \le \sum_{(s,a):\, d_h^\mu(s,a) > 0} d_h^{\pi^\star}(s,a) \left(1 + \frac{8\iota}{d_h^\mu(s,a)} \right) \le 1 + 8SAC_\eta^\star \iota,$$

which proves the third claim of Lemma B.4.

In the sum in equation (6), the $k=1$ contribution is zero. For $k \ge 2$, (3) with $m = k-1$ gives

$$\sum_{s,a} d_h^{\pi^\star}(s,a) \frac{\mathbf{1}\{N_h^k(s,a) \ge 1\}}{N_h^k(s,a) \vee 1} = \sum_{s,a} \frac{d_h^{\pi^\star}(s,a)}{d_h^\mu(s,a)} \frac{d_h^\mu(s,a) \mathbf{1}\{N_h^k(s,a) \ge 1\}}{N_h^k(s,a) \vee 1} \le \frac{8SAC_\eta^\star \iota}{k-1}.$$

Summing over k and using $\sum_{j=1}^{K-1} 1/j \le 1 + \log K$ proves the fourth claim of Lemma B.4.

Finally, (3) with $m = K$ gives $d_h^\mu(s,a)/(N_h^{K+1}(s,a) \vee 1) \le 8\iota/K$. Hence

$$\sum_{s,a} d_h^{\pi^\star}(s,a)\mathbf{1}\{N_h^{K+1}(s,a) = 0\} \le \frac{8\iota}{K}\sum_{s,a}\frac{d_h^{\pi^\star}(s,a)}{d_h^\mu(s,a)} \le \frac{8SAC_\eta^\star\iota}{K},$$
$$\sum_{s,a} d_h^{\pi^\star}(s,a)\frac{\mathbf{1}\{N_h^{K+1}(s,a) \ge 1\}}{N_h^{K+1}(s,a) \vee 1} \le \frac{8\iota}{K}\sum_{s,a}\frac{d_h^{\pi^\star}(s,a)}{d_h^\mu(s,a)} \le \frac{8SAC_\eta^\star\iota}{K}.$$

This proves the final two claims of Lemma B.4. □

The following concentration bound is an immediate consequence of Lemma B.6 of Shi et al. (2022), applied with $C_d = 1$, $C_w = B_h$, and $W_{h+1}^k(s,a) = U_h^k(s,a)\mathbf{1}$, where $\mathbf{1}$ denotes the all-ones state vector, together with $(U_h^k(s,a))^2 \le B_h U_h^k(s,a)$. The logarithmic factors are absorbed into ι.

Lemma B.5. *Suppose Assumption 2.2 holds. Fix nonnegative $\mathcal{D}_{<k}$-measurable functions $U_h^k : \mathcal{S}\times\mathcal{A} \to [0, B_h]$. For any $\delta \in (0,1)$, let $\iota := \log\frac{SAT}{\delta}$. With probability at least $1-\delta$, simultaneously for every $h \in [H]$,*

$$\left|\sum_{k=1}^{K}\frac{d_h^{\pi^\star}(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)}U_h^k(s_h^k,a_h^k) - \sum_{k=1}^{K}\sum_{s,a} d_h^{\pi^\star}(s,a)U_h^k(s,a)\right|$$
$$\le c\sqrt{C_\eta^\star B_h\iota\sum_{k=1}^{K}\sum_{s,a} d_h^{\pi^\star}(s,a)U_h^k(s,a) + cC_\eta^\star B_h\iota}.$$

C Proofs for Algorithm 1

Throughout, constant factors in logarithmic confidence terms are absorbed into universal constants. For notational convenience, we define $\Delta_h^k(s,a) := Q_h^k(s,a) - Q_h^\star(s,a)$.

C.1 Lemmas for Theorem 3.1

Lemma C.1. *For every $k \in [K+1]$, $h \in [H]$, and $(s,a) \in \mathcal{S}\times\mathcal{A}$,*

$$0 \le Q_h^k(s,a) \le L_h, \qquad 0 \le V_h^k(s) \le L_h.$$

Proof. The claim holds at initialization. Suppose it holds before episode k. At a stage-h update,

$$0 \le r_h^k + V_{h+1}^k(s_{h+1}^k) \le 1 + L_{h+1} = L_h.$$

Thus the updated Q-value is a convex combination of two values in $[0, L_h]$, while all other Q-values are unchanged. Since $F_{h,s}(q)$ lies between $\min_a q(a)$ and $\max_a q(a)$, the updated value also lies in $[0, L_h]$. Induction over episodes proves the claim. □

Lemma C.2. *For every $k \in [K+1]$, $h \in [H]$, and $s \in \mathcal{S}$,*

$$(V_h^k(s) - V_h^\star(s))^2 \le \gamma_h e_h^k(s). \tag{9}$$

Moreover, the output policy in Algorithm 1 satisfies

$$\mathrm{SubOpt}(\widehat{\pi}) \le \frac{\eta}{2}\sum_{h=1}^{H}\Gamma_{1:h}\mathbb{E}_{(s,a)\sim d_h^{\pi^\star}}\left[(Q_h^{K+1}(s,a) - Q_h^\star(s,a))^2\right]. \tag{10}$$

Proof. Fix $k \in [K+1]$, $h \in [H]$, and $s \in \mathcal{S}$. By Lemma C.1 and Proposition 2.1, for every $a \in \mathcal{A}$,

$$Q_h^k(s,a), Q_h^\star(s,a) \in [0, L_h], \qquad -L_h \le \Delta_h^k(s,a) \le L_h.$$

For $t \in [0,1]$, set $Q^k_{h,t}(s,\cdot) := Q^\star_h(s,\cdot) + t\Delta^k_h(s,\cdot)$. Using the Gibbs formula and $\pi^\star_h(\cdot \mid s) = G_{h,s}(Q^\star_h(s,\cdot))$, for every action in the support of $\pi^{\mathrm{ref}}_h(\cdot \mid s)$,

$$\begin{aligned} G_{h,s}(Q^k_{h,t}(s,\cdot))(a) &= \frac{\pi^\star_h(a \mid s)e^{\eta t\Delta^k_h(s,a)}}{\sum_{b\in\mathcal{A}} \pi^\star_h(b \mid s)e^{\eta t\Delta^k_h(s,b)}} \\ &\le \pi^\star_h(a \mid s)e^{2\eta t L_h} \le \gamma_h \pi^\star_h(a \mid s). \end{aligned} \tag{11}$$

The same inequality is trivial outside the support, where both probabilities are zero. Since $\nabla F_{h,s} = G_{h,s}$, the fundamental theorem of calculus gives

$$\begin{aligned} (V^k_h(s) - V^\star_h(s))^2 &= \left(\int_0^1 \left\langle G_{h,s}(Q^k_{h,t}(s,\cdot)), \Delta^k_h(s,\cdot)\right\rangle_{\mathcal{A}} dt\right)^2 \\ &\le \int_0^1 \sum_{a\in\mathcal{A}} G_{h,s}(Q^k_{h,t}(s,\cdot))(a)(\Delta^k_h(s,a))^2\, dt \\ &\le \gamma_h \sum_{a\in\mathcal{A}} \pi^\star_h(a \mid s)(\Delta^k_h(s,a))^2 = \gamma_h e^k_h(s), \end{aligned}$$

where the first inequality follows from Jensen's inequality, applied first to the integral and then to the action distribution, and the second follows from (11). This proves the first claim of Lemma C.2.

We next bound the KL divergence between the corresponding Gibbs policies. Define

$$\phi(t) := \log \sum_{b\in\mathcal{A}} \pi^{\mathrm{ref}}_h(b \mid s)e^{\eta Q^k_{h,1-t}(s,b)}, \qquad t \in [0,1].$$

The Gibbs formulas imply

$$\mathrm{KL}\big(G_{h,s}(Q^k_h(s,\cdot))\big\|\pi^\star_h(\cdot \mid s)\big) = \phi(1) - \phi(0) - \phi'(0),$$

and differentiating $\phi(t)$ twice yields

$$\phi''(t) = \eta^2 \operatorname{Var}_{a\sim G_{h,s}(Q^k_{h,1-t}(s,\cdot))}(\Delta^k_h(s,a)).$$

Therefore, Taylor's theorem with integral remainder, together with $\mathrm{Var}(X) \le \mathbb{E}[X^2]$ and (11), yields

$$\begin{aligned} \mathrm{KL}\big(G_{h,s}(Q^k_h(s,\cdot))\big\|\pi^\star_h(\cdot \mid s)\big) &= \eta^2 \int_0^1 (1-t) \operatorname{Var}_{a\sim G_{h,s}(Q^k_{h,1-t}(s,\cdot))}(\Delta^k_h(s,a))\, dt \\ &\le \eta^2\gamma_h \int_0^1 (1-t)\, dt \sum_{a\in\mathcal{A}} \pi^\star_h(a \mid s)(\Delta^k_h(s,a))^2 = \frac{\eta^2\gamma_h}{2} e^k_h(s). \end{aligned} \tag{12}$$

We now turn to the output policy. Taking $k = K+1$ and $t = 1$ in (11) gives $\widehat{\pi}_h(a \mid s) \le \gamma_h \pi^\star_h(a \mid s)$. Since $d^{\widehat{\pi}}_1 = d^{\pi^\star}_1 = \rho$, induction over h gives

$$\begin{aligned} d^{\widehat{\pi}}_{h+1}(s') &= \sum_{s,a} d^{\widehat{\pi}}_h(s)\widehat{\pi}_h(a \mid s)\mathbb{P}_h(s' \mid s,a) \\ &\le \Gamma_{1:h-1}\gamma_h \sum_{s,a} d^{\pi^\star}_h(s)\pi^\star_h(a \mid s)\mathbb{P}_h(s' \mid s,a) = \Gamma_{1:h} d^{\pi^\star}_{h+1}(s'). \end{aligned} \tag{13}$$

Thus $d^{\widehat{\pi}}_h(s) \le \Gamma_{1:h-1} d^{\pi^\star}_h(s)$ for every h and s.

It remains to convert the pointwise KL bound into policy suboptimality. For any policy π with finite regularized value, the Gibbs formula for $\pi^\star$ gives

$$\log \frac{\pi^\star_h(a \mid s)}{\pi^{\mathrm{ref}}_h(a \mid s)} = \eta(Q^\star_h(s,a) - V^\star_h(s))$$

on the common support. Combining this identity with Bellman consistency yields

$$V_h^\star(s) - V_h^\pi(s) = \frac{1}{\eta}\mathrm{KL}(\pi_h(\cdot \mid s)\|\pi_h^\star(\cdot \mid s)) + \sum_a \pi_h(a \mid s)(\mathbb{P}_h(V_{h+1}^\star - V_{h+1}^\pi))(s,a).$$

Taking expectation with respect to $s \sim d_h^\pi$ and summing over h, the next-stage value terms telescope, yielding

$$J_\eta(\pi^\star) - J_\eta(\pi) = \frac{1}{\eta}\sum_{h=1}^H \mathbb{E}_{s\sim d_h^\pi}\left[\mathrm{KL}(\pi_h(\cdot \mid s)\|\pi_h^\star(\cdot \mid s))\right].$$

Taking $\pi = \widehat{\pi}$, using (12) and (13), and then expanding e_h^{K+1} gives

$$\begin{aligned}
\mathrm{SubOpt}(\widehat{\pi}) &\le \frac{\eta}{2}\sum_{h=1}^H \gamma_h \sum_s d_h^{\widehat{\pi}}(s) e_h^{K+1}(s) \\
&\le \frac{\eta}{2}\sum_{h=1}^H \Gamma_{1:h} \sum_s d_h^{\pi^\star}(s) e_h^{K+1}(s) \\
&= \frac{\eta}{2}\sum_{h=1}^H \Gamma_{1:h}\mathbb{E}_{(s,a)\sim d_h^{\pi^\star}}\left[(Q_h^{K+1}(s,a) - Q_h^\star(s,a))^2\right],
\end{aligned}$$

which proves the second claim of Lemma C.2. □

We next combine Lemma B.3 with the KL value-error bound in Lemma C.2 to derive a pointwise recursion for the squared Q-estimation error. This recursion will be used to control both the cumulative squared error and the error of the final iterate.

Lemma C.3. *For any $\delta \in (0,1)$, let $\iota := \log\frac{SAT}{\delta}$. With probability at least $1-\delta$, simultaneously for every $k \in [K+1]$, $h \in [H]$, and $(s,a) \in \mathcal{S}\times\mathcal{A}$,*

$$\begin{aligned}
(\Delta_h^k(s,a))^2 \le L_h^2 \mathbf{1}\{N_h^k(s,a) = 0\} + c_0 H^2 L_{h+1}^2 \iota^2 \frac{\mathbf{1}\{N_h^k(s,a) \ge 1\}}{N_h^k(s,a) \vee 1} \\
+ \left(1 + \frac{1}{H}\right)\gamma_{h+1} \sum_{n=1}^{N_h^k(s,a)} \alpha_n^{N_h^k(s,a)} (\mathbb{P}_h e_{h+1}^{k_{h,n}(s,a)})(s,a),
\end{aligned} \tag{14}$$

where $c_0 > 0$ is a universal constant.

Proof. By Lemma C.1, V_{h+1}^k satisfies $0 \le V_{h+1}^k \le L_{h+1}$, and by the update order of Algorithm 1, V_{h+1}^k is $\mathcal{D}_{<k}$-measurable. Hence Lemma B.3 applies. Work on its event of probability at least $1-\delta$. Fix (k,h,s,a), write $N := N_h^k(s,a)$, and suppress the dependence of $k_{h,n}$ on (s,a). If $N = 0$, then $Q_h^k(s,a) = 0$ and $(\Delta_h^k(s,a))^2 \le L_h^2$ by (1).

Suppose $N \ge 1$. Unrolling the Q-learning update and subtracting the optimal Bellman equation with the same coefficients α_n^N gives

$$\Delta_h^k(s,a) = \sum_{n=1}^N \alpha_n^N (\widehat{\mathbb{P}}_h^{k_{h,n}} - \mathbb{P}_{h,s,a})V_{h+1}^{k_{h,n}} + \sum_{n=1}^N \alpha_n^N \mathbb{P}_{h,s,a}(V_{h+1}^{k_{h,n}} - V_{h+1}^\star),$$

since $\alpha_0^N = 0$. Applying $(x+y)^2 \le (H+1)x^2 + (1+1/H)y^2$ and Lemma B.3 yields

$$\begin{aligned}
(\Delta_h^k(s,a))^2 &\le \frac{c_0 H^2 L_{h+1}^2 \iota^2}{N} + \left(1+\frac{1}{H}\right)\left(\sum_{n=1}^N \alpha_n^N \mathbb{P}_{h,s,a}(V_{h+1}^{k_{h,n}} - V_{h+1}^\star)\right)^2 \\
&\le \frac{c_0 H^2 L_{h+1}^2 \iota^2}{N} + \left(1+\frac{1}{H}\right)\sum_{n=1}^N \alpha_n^N \mathbb{P}_{h,s,a}\left[(V_{h+1}^{k_{h,n}} - V_{h+1}^\star)^2\right] \\
&\le \frac{c_0 H^2 L_{h+1}^2 \iota^2}{N} + \left(1+\frac{1}{H}\right)\gamma_{h+1}\sum_{n=1}^N \alpha_n^N (\mathbb{P}_h e_{h+1}^{k_{h,n}})(s,a).
\end{aligned}$$

The second inequality follows from Jensen's inequality, first with respect to the weights $\{\alpha_n^N\}$ and then with respect to $\mathbb{P}_h(\cdot \mid s,a)$, the last follows from (9) at stage $h+1$. Together with the zero-count case, this proves Lemma C.3. □

Lemma C.4. *Suppose Assumption 2.2 holds and $SAT > 1$. For any $\delta \in (0,1)$, let $\iota := \log \frac{SAT}{\delta}$. Then, with probability at least $1-\delta$, simultaneously for every $h \in [H]$,*

$$\sum_{k=1}^{K}\sum_{s} d_h^{\pi^\star}(s) e_h^k(s) \le \left(1+\frac{5}{H}\right)\gamma_{h+1}\sum_{k=1}^{K}\sum_{s} d_{h+1}^{\pi^\star}(s) e_{h+1}^k(s) + c_0 SAC_\eta^\star \gamma_{h+1} H^2 L_h^2 \iota^3 (1+\log K). \tag{15}$$

Consequently,

$$\sum_{k=1}^{K} \mathbb{E}_{(s,a)\sim d_h^{\pi^\star}}\left[(Q_h^k(s,a) - Q_h^\star(s,a))^2\right] \le c_1 SAC_\eta^\star \Gamma_{h+1:H} H^2 L_h^3 \iota^3 (1+\log K), \tag{16}$$

where $c_0, c_1 > 0$ are universal constants.

Proof. Invoke Lemmas C.3 and B.4 with failure probability $\delta/4$ each, and invoke Lemma B.5 twice with failure probability $\delta/4$, first with

$$U_h^k(s,a) := \sum_{n=1}^{N_h^k(s,a)} \alpha_n^{N_h^k(s,a)} (\mathbb{P}_h e_{h+1}^{k_{h,n}(s,a)})(s,a), \tag{17}$$

and then with

$$U_h^k(s,a) := (\mathbb{P}_h e_{h+1}^k)(s,a). \tag{18}$$

In both applications take $B_h = L_{h+1}^2$. By Lemma C.1 and Proposition 2.1, $0 \le e_{h+1}^k(s) \le L_{h+1}^2$. Moreover, by the update order in Algorithm 1, Q^k, and hence e^k, are $\mathcal{D}_{<k}$-measurable. Since $k_{h,n}(s,a) < k$ for every $1 \le n \le N_h^k(s,a)$, the historical test function in (17) is also $\mathcal{D}_{<k}$-measurable. Thus both choices satisfy the hypotheses of Lemma B.5. By a union bound, all four events hold simultaneously with probability at least $1-\delta$.

Summing (14) over $k \in [K]$ and averaging with respect to $d_h^{\pi^\star}$ gives

$$\begin{aligned}\sum_{k=1}^{K}\sum_{s} d_h^{\pi^\star}(s) e_h^k(s) \le{}& L_h^2 \sum_{k=1}^{K}\sum_{s,a} d_h^{\pi^\star}(s,a)\mathbf{1}\{N_h^k(s,a)=0\} \\ &+ c_0 H^2 L_{h+1}^2 \iota^2 \sum_{k=1}^{K}\sum_{s,a} d_h^{\pi^\star}(s,a)\frac{\mathbf{1}\{N_h^k(s,a)\ge 1\}}{N_h^k(s,a)\vee 1} \\ &+ \left(1+\frac{1}{H}\right)\gamma_{h+1}\underbrace{\sum_{k=1}^{K}\sum_{s,a} d_h^{\pi^\star}(s,a)\sum_{n=1}^{N_h^k(s,a)} \alpha_n^{N_h^k(s,a)} (\mathbb{P}_h e_{h+1}^{k_{h,n}(s,a)})(s,a)}_{\text{(I)}}.\end{aligned} \tag{19}$$

By (5) and (6), we have

$$\begin{aligned}&L_h^2 \sum_{k=1}^{K}\sum_{s,a} d_h^{\pi^\star}(s,a)\mathbf{1}\{N_h^k(s,a)=0\} + c_0 H^2 L_{h+1}^2 \iota^2 \sum_{k=1}^{K}\sum_{s,a} d_h^{\pi^\star}(s,a)\frac{\mathbf{1}\{N_h^k(s,a)\ge 1\}}{N_h^k(s,a)\vee 1} \\ &\quad\le cSAC_\eta^\star H^2 L_h^2 \iota^3 (1+\log K).\end{aligned} \tag{20}$$

It remains to bound (I). Applying Lemma B.5 with the test function in (17) gives

$$\begin{aligned}\text{(I)} &\le \underbrace{\sum_{k=1}^{K}\frac{d_h^{\pi^\star}(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)}\sum_{n=1}^{N_h^k(s_h^k,a_h^k)} \alpha_n^{N_h^k(s_h^k,a_h^k)} (\mathbb{P}_h e_{h+1}^{k_{h,n}(s_h^k,a_h^k)})(s_h^k,a_h^k)}_{\text{(II)}} + c\sqrt{C_\eta^\star L_{h+1}^2 \iota\,\text{(I)}} + cC_\eta^\star L_{h+1}^2\iota \\ &\overset{(i)}{\le} \left(1+\frac{1}{8H}\right)\text{(II)} + cC_\eta^\star H L_{h+1}^2 \iota.\end{aligned} \tag{21}$$

Here (i) uses $c\sqrt{xy} \le \frac{x}{16H} + 4c^2Hy$ with $x = \text{(I)}$ and $y = C_\eta^\star L_{h+1}^2 \iota$. After moving $x/(16H)$ to the left, we use $(1 - 1/(16H))^{-1} \le 1 + 1/(8H)$ and absorb numerical factors into the universal constant.

We next reindex (II). For each state-action pair (s,a), the m-th visit occurs in episode $k_{h,m}(s,a)$ and satisfies $N_h^{k_{h,m}(s,a)}(s,a) = m-1$. Hence

$$\begin{aligned}
\text{(II)} &= \sum_{s,a} \frac{d_h^{\pi^\star}(s,a)}{d_h^\mu(s,a)} \sum_{m=1}^{N_h^{K+1}(s,a)} \sum_{n=1}^{m-1} \alpha_n^{m-1} (\mathbb{P}_h e_{h+1}^{k_{h,n}(s,a)})(s,a) \\
&= \sum_{s,a} \frac{d_h^{\pi^\star}(s,a)}{d_h^\mu(s,a)} \sum_{n=1}^{N_h^{K+1}(s,a)-1} \left(\sum_{m=n+1}^{N_h^{K+1}(s,a)} \alpha_n^{m-1} \right) (\mathbb{P}_h e_{h+1}^{k_{h,n}(s,a)})(s,a) \\
&\le \left(1 + \frac{1}{H}\right) \sum_{s,a} \frac{d_h^{\pi^\star}(s,a)}{d_h^\mu(s,a)} \sum_{n=1}^{N_h^{K+1}(s,a)} (\mathbb{P}_h e_{h+1}^{k_{h,n}(s,a)})(s,a) \\
&= \left(1 + \frac{1}{H}\right) \underbrace{\sum_{k=1}^{K} \frac{d_h^{\pi^\star}(s_h^k, a_h^k)}{d_h^\mu(s_h^k, a_h^k)} (\mathbb{P}_h e_{h+1}^k)(s_h^k, a_h^k)}_{\text{(III)}} . \qquad (22)
\end{aligned}$$

The inequality in (22) uses

$$\sum_{m=n+1}^{N_h^{K+1}(s,a)} \alpha_n^{m-1} \le \sum_{M=n}^{\infty} \alpha_n^M = 1 + \frac{1}{H},$$

and adds the nonnegative term corresponding to $n = N_h^{K+1}(s,a)$. It remains to control (III). Applying Lemma B.5 with (18) and the occupancy recursion under $\pi^\star$ gives

$$\begin{aligned}
\text{(III)} &\le \sum_{k=1}^{K} \sum_{s,a} d_h^{\pi^\star}(s,a)(\mathbb{P}_h e_{h+1}^k)(s,a) + c\sqrt{C_\eta^\star L_{h+1}^2 \iota \sum_{k=1}^{K} \sum_{s,a} d_h^{\pi^\star}(s,a)(\mathbb{P}_h e_{h+1}^k)(s,a)} + cC_\eta^\star L_{h+1}^2 \iota \\
&= \sum_{k=1}^{K} \sum_{s} d_{h+1}^{\pi^\star}(s) e_{h+1}^k(s) + c\sqrt{C_\eta^\star L_{h+1}^2 \iota \sum_{k=1}^{K} \sum_{s} d_{h+1}^{\pi^\star}(s) e_{h+1}^k(s)} + cC_\eta^\star L_{h+1}^2 \iota \\
&\overset{(ii)}{\le} \left(1 + \frac{1}{16H}\right) \sum_{k=1}^{K} \sum_{s} d_{h+1}^{\pi^\star}(s) e_{h+1}^k(s) + cC_\eta^\star H L_{h+1}^2 \iota . \qquad (23)
\end{aligned}$$

Here (ii) uses

$$c\sqrt{xy} \le \frac{x}{16H} + 4c^2 Hy,$$

with

$$x = \sum_{k=1}^{K} \sum_{s} d_{h+1}^{\pi^\star}(s) e_{h+1}^k(s), \qquad y = C_\eta^\star L_{h+1}^2 \iota .$$

Combining (21), (22), and (23) gives

$$\begin{aligned}
\text{(I)} &\le \left(1 + \frac{1}{8H}\right)\left(1 + \frac{1}{H}\right)\left(1 + \frac{1}{16H}\right) \sum_{k=1}^{K} \sum_{s} d_{h+1}^{\pi^\star}(s) e_{h+1}^k(s) + cC_\eta^\star H L_{h+1}^2 \iota \\
&\le \left(1 + \frac{2}{H}\right) \sum_{k=1}^{K} \sum_{s} d_{h+1}^{\pi^\star}(s) e_{h+1}^k(s) + cC_\eta^\star H L_{h+1}^2 \iota . \qquad (24)
\end{aligned}$$

Combining (19), (20), and (24) yields

$$\sum_{k=1}^{K}\sum_{s} d_h^{\pi^\star}(s)e_h^k(s) \le cSAC_\eta^\star H^2 L_h^2 \iota^3(1+\log K)+ \left(1+\frac{1}{H}\right)\gamma_{h+1}\left[\left(1+\frac{2}{H}\right)\sum_{k=1}^{K}\sum_{s} d_{h+1}^{\pi^\star}(s)e_{h+1}^k(s) + cC_\eta^\star H L_{h+1}^2\iota\right] \le \left(1+\frac{5}{H}\right)\gamma_{h+1}\sum_{k=1}^{K}\sum_{s} d_{h+1}^{\pi^\star}(s)e_{h+1}^k(s) + c_0 SAC_\eta^\star\gamma_{h+1}H^2L_h^2\iota^3(1+\log K).$$

Here we used

$$\left(1+\frac{1}{H}\right)\left(1+\frac{2}{H}\right) = 1+\frac{3}{H}+\frac{2}{H^2} \le 1+\frac{5}{H},$$

together with $\gamma_{h+1} \ge 1$, $L_{h+1} \le L_h$, $SA \ge 1$, $H \ge 1$, $\iota \ge \log 2$, and $1+\log K \ge 1$, absorbing lower-order terms into the universal constant. This proves the first claim of Lemma C.4.

Iterating (15) backward, with the stage-$H+1$ term equal to zero, gives

$$\sum_{k=1}^{K}\sum_{s} d_h^{\pi^\star}(s)e_h^k(s) \le c_0 SAC_\eta^\star H^2\iota^3(1+\log K)\sum_{t=h}^{H}\left(1+\frac{5}{H}\right)^{t-h}\left(\prod_{j=h+1}^{t+1}\gamma_j\right)L_t^2 \le cSAC_\eta^\star\Gamma_{h+1:H}H^2\iota^3(1+\log K)\sum_{t=h}^{H}L_t^2 \le c_1 SAC_\eta^\star\Gamma_{h+1:H}H^2L_h^3\iota^3(1+\log K),$$

where we used

$$\left(1+\frac{5}{H}\right)^{t-h} \le \left(1+\frac{5}{H}\right)^{H} \le e^5, \qquad \prod_{j=h+1}^{t+1}\gamma_j \le \prod_{j=h+1}^{H}\gamma_j = \Gamma_{h+1:H},$$

and

$$\sum_{t=h}^{H}L_t^2 = \sum_{\ell=1}^{L_h}\ell^2 \le L_h^3.$$

By the definition of e_h^k, this proves the second claim of Lemma C.4. □

Lemma C.5. *Suppose Assumption 2.2 holds. For any $\delta \in (0,1)$, let $\iota := \log\frac{SAT}{\delta}$. Then, with probability at least $1-\delta$, simultaneously for every $h \in [H]$,*

$$\mathbb{E}_{(s,a)\sim d_h^{\pi^\star}}\left[(Q_h^{K+1}(s,a)-Q_h^\star(s,a))^2\right] \le \frac{c_0 SAC_\eta^\star\gamma_{h+1}H^2L_h^2\iota^3}{K} + \frac{c_0 H\iota\gamma_{h+1}}{K}\sum_{k=1}^{K}\mathbb{E}_{s\sim d_{h+1}^{\pi^\star}}\left[e_{h+1}^k(s)\right], \tag{25}$$

where $c_0 > 0$ is a universal constant, and the sum on the right is zero when $h = H$.

Proof. If $SAT = 1$, the single Q-update is exact, so the left-hand side is zero. Hence suppose $SAT > 1$. Invoke Lemmas C.3, B.4, and B.5 with failure probability $\delta/3$ respectively, for Lemma B.5, take $U_h^k(s,a) = (\mathbb{P}_h e_{h+1}^k)(s,a)$ and $B_h = L_{h+1}^2$. The range bounds ensure $0 \le U_h^k(s,a) \le L_{h+1}^2$, and U_h^k is $\mathcal{D}_{<k}$-measurable. By a union bound, all three events hold simultaneously with probability at least $1-\delta$.

Fix $h \in [H]$. For every state-action pair with $d_h^{\pi^\star}(s,a) > 0$ and $N_h^{K+1}(s,a) \ge 1$, Assumption 2.2 gives $d_h^\mu(s,a) > 0$, and Lemma B.1 together with (3) at $m = K$ and $\log(3SAT/\delta) \le 3\iota$ gives

$$\max_n \alpha_n^{N_h^{K+1}(s,a)} \le \frac{2H}{N_h^{K+1}(s,a)} \le \frac{48H\iota}{Kd_h^\mu(s,a)}. \tag{26}$$

Averaging (14) at $k = K+1$ with respect to $d_h^{\pi^\star}$ and using (7)–(8), we obtain

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim d_h^{\pi^\star}}\left[(Q_h^{K+1}(s,a) - Q_h^\star(s,a))^2\right] \\
&\quad \le \frac{cSAC_\eta^\star H^2 L_h^2 \iota^3}{K} + \left(1+\frac{1}{H}\right)\gamma_{h+1}\sum_{s,a} d_h^{\pi^\star}(s,a) \sum_{n=1}^{N_h^{K+1}(s,a)} \alpha_n^{N_h^{K+1}(s,a)} (\mathbb{P}_h e_{h+1}^{k_{h,n}(s,a)})(s,a) \\
&\quad \overset{(i)}{\le} \frac{cSAC_\eta^\star H^2 L_h^2 \iota^3}{K} + \frac{48H\iota}{K}\left(1+\frac{1}{H}\right)\gamma_{h+1} \underbrace{\sum_{k=1}^{K} \frac{d_h^{\pi^\star}(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)} (\mathbb{P}_h e_{h+1}^k)(s_h^k,a_h^k)}_{\text{(I)}}. \qquad (27)
\end{aligned}$$

Here (i) follows from (26) after reindexing the pairs $((s,a),n)$ by their visit episodes $k_{h,n}(s,a)$, the factor $d_h^{\pi^\star}(s,a)/d_h^\mu(s,a)$ then yields (I).

We next control (I). Applying Lemma B.5 with $U_h^k(s,a) = (\mathbb{P}_h e_{h+1}^k)(s,a)$ and $B_h = L_{h+1}^2$, and then using the occupancy recursion under $\pi^\star$,

$$\text{(I)} \le \sum_{k=1}^{K}\sum_{s,a} d_h^{\pi^\star}(s,a)(\mathbb{P}_h e_{h+1}^k)(s,a) + c\sqrt{C_\eta^\star L_{h+1}^2 \iota \sum_{k=1}^{K}\sum_{s,a} d_h^{\pi^\star}(s,a)(\mathbb{P}_h e_{h+1}^k)(s,a)} + cC_\eta^\star L_{h+1}^2\iota \qquad (28)$$

$$\begin{aligned}
&= \sum_{k=1}^{K}\sum_{s} d_{h+1}^{\pi^\star}(s) e_{h+1}^k(s) + c\sqrt{C_\eta^\star L_{h+1}^2 \iota \sum_{k=1}^{K}\sum_{s} d_{h+1}^{\pi^\star}(s) e_{h+1}^k(s)} + cC_\eta^\star L_{h+1}^2\iota \\
&\overset{(ii)}{\le} c\sum_{k=1}^{K}\sum_{s} d_{h+1}^{\pi^\star}(s) e_{h+1}^k(s) + cC_\eta^\star L_{h+1}^2\iota. \qquad (29)
\end{aligned}$$

Here (ii) uses $\sqrt{xy} \le (x+y)/2$ with

$$x = \sum_{k=1}^{K}\sum_{s} d_{h+1}^{\pi^\star}(s) e_{h+1}^k(s), \qquad y = C_\eta^\star L_{h+1}^2\iota,$$

after adjusting the universal constant c. Combining (27) and (29) gives

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim d_h^{\pi^\star}}\left[(Q_h^{K+1}(s,a) - Q_h^\star(s,a))^2\right] \\
&\quad \le \frac{cSAC_\eta^\star H^2 L_h^2 \iota^3}{K} + \frac{cH\iota\gamma_{h+1}}{K}\sum_{k=1}^{K}\sum_{s} d_{h+1}^{\pi^\star}(s) e_{h+1}^k(s) + \frac{cC_\eta^\star \gamma_{h+1} H L_{h+1}^2 \iota^2}{K} \\
&\quad \le \frac{c_0 SAC_\eta^\star \gamma_{h+1} H^2 L_h^2 \iota^3}{K} + \frac{c_0 H\iota\gamma_{h+1}}{K}\sum_{k=1}^{K} \mathbb{E}_{s\sim d_{h+1}^{\pi^\star}}\left[e_{h+1}^k(s)\right].
\end{aligned}$$

Here we used $L_{h+1} \le L_h$, $SA \ge 1$, $\iota \ge \log 2$, $\gamma_{h+1} \ge 1$, and the definition of e_{h+1}^k. This proves Lemma C.5. When $h = H$, the stage-$H+1$ terms are zero. □

C.2 Proof of Theorem 3.1

Proof. If $SAT = 1$, the output is the unique policy, so $\mathrm{SubOpt}(\widehat{\pi}) = 0$. Hence suppose $SAT > 1$. Apply Lemmas C.5 and C.4 with failure probability $\delta/2$ each. With probability at least $1-\delta$, for every $h \in [H]$,

$$\begin{aligned}
\mathbb{E}_{(s,a)\sim d_h^{\pi^\star}}\left[(Q_h^{K+1}(s,a) - Q_h^\star(s,a))^2\right] &\le \frac{cSAC_\eta^\star \gamma_{h+1} H^2 L_h^2 \iota^3}{K} + \frac{cH\iota\gamma_{h+1}}{K}\sum_{k=1}^{K}\sum_{s} d_{h+1}^{\pi^\star}(s) e_{h+1}^k(s) \\
&\le c\frac{SAC_\eta^\star}{K}\Gamma_{h+1:H} H^3 L_h^3 \iota^4 (1+\log K),
\end{aligned}$$

where the second inequality uses (16) at stage $h+1$, when $h = H$, $e^k_{H+1} \equiv 0$. Substituting this estimate into (10) gives

$$\begin{aligned}\mathrm{SubOpt}(\widehat{\pi}) &\le c\frac{\eta SAC^\star_\eta H^3\iota^4(1+\log K)}{K}\sum_{h=1}^{H}\Gamma_{1:h}\Gamma_{h+1:H}L_h^3\\ &= c\frac{\eta e^{\eta H(H+1)}SAC^\star_\eta H^3\iota^4(1+\log K)}{K}\sum_{h=1}^{H}L_h^3\\ &\le c\frac{\eta e^{\eta H(H+1)}H^7SAC^\star_\eta}{K}\iota^4(1+\log K).\end{aligned}$$

Here $\Gamma_{1:h}\Gamma_{h+1:H} = \Gamma_{1:H} = e^{\eta H(H+1)}$ and $\sum_{h=1}^{H}L_h^3 = (H(H+1)/2)^2 \le H^4$. This proves Theorem 3.1. □

D Proofs for Algorithm 2

Throughout this appendix, fix the confidence level $\delta \in (0,1)$ used in Algorithm 2 and let $\iota := \log\frac{SAT}{\delta}$. All iterate-dependent quantities refer to this fixed run of KLPQ. We use the shorthand $\Delta^k_h(s,a) := Q^k_h(s,a) - Q^\star_h(s,a)$, where Q^k_h now denotes the KLPQ iterates, and retain the definition of e^k_h from Section 5.1. We also retain $\mathcal{D}_{<k}$ and $\widehat{\mathbb{P}}^k_h$ from Appendix B. Constant-factor changes in concentration thresholds are absorbed into universal constants, throughout, the algorithm is run with the original confidence level.

D.1 Lemmas for Theorem 4.1

Lemma D.1. *For every $k \in [K+1]$, $h \in [H]$, and $(s,a) \in \mathcal{S}\times\mathcal{A}$,*

$$\widetilde{Q}^k_h(s,a) \le L_h, \qquad 0 \le Q^k_h(s,a) \le L_h, \qquad 0 \le V^k_h(s) \le L_h. \tag{30}$$

Moreover, $Q^{k+1}_h \ge Q^k_h$ and $V^{k+1}_h \ge V^k_h$ coordinatewise for every $k \in [K]$ and $h \in [H]$, and

$$Q^k_h(s,a) = \max_{1\le j\le k}\widetilde{Q}^j_h(s,a), \qquad V^k_h(s) = F_{h,s}(Q^k_h(s,\cdot)).$$

Proof. We prove the stated properties by induction over episodes.

For the base case $k = 1$, all claims follow directly from the initialization, and $V^1_h(s) = F_{h,s}(Q^1_h(s,\cdot))$ for every $(h,s) \in [H]\times\mathcal{S}$. For the induction hypothesis, suppose that the range bounds in (30), the invariant $V^k_h(s) = F_{h,s}(Q^k_h(s,\cdot))$, and the identity $Q^k_h(s,a) = \max_{1\le j\le k}\widetilde{Q}^j_h(s,a)$ hold at the beginning of episode k.

For the induction step from k to $k+1$, the algorithm carries Q^k_h, V^k_h, and $\widetilde{Q}^k_h$ forward except at the visited coordinates. At a stage-h update, $r^k_h + V^k_{h+1}(s^k_{h+1}) - b_{h,n} \le 1 + L_{h+1} = L_h$. Hence the update of $\widetilde{Q}_h$ remains at most L_h. The max update in Line 12 therefore ensures that Q^{k+1}_h remains in $[0, L_h]$, preserves coordinatewise monotonicity, and yields $Q^{k+1}_h(s,a) = \max_{1\le j\le k+1}\widetilde{Q}^j_h(s,a)$. Since $F_{h,s}$ is coordinatewise nondecreasing and lies between the minimum and maximum coordinates of its argument, the recomputed value also lies in $[0, L_h]$ and satisfies $V^{k+1}_h \ge V^k_h$ coordinatewise. The invariant $V^{k+1}_h(s) = F_{h,s}(Q^{k+1}_h(s,\cdot))$ is preserved because the algorithm carries Q^k_h and V^k_h forward together and recomputes the value at every updated state. This completes the induction over k. □

We now introduce a concentration event for later use. Notably, Q^k and V^k before episode k depend only on $\tau^1,\dots,\tau^{k-1}$, so Q^k_h, V^k_h, and e^k_h are $\mathcal{D}_{<k}$-measurable. The deterministic bounds (30) and (1) give,

$$0 \le e^k_h(s) \le L_h^2, \qquad 0 \le (\mathbb{P}_h e^k_{h+1})(s,a) \le L^2_{h+1},$$

$$0 \le \sum_{n=1}^{N^k_h(s,a)} \alpha_n^{N^k_h(s,a)}(\mathbb{P}_h e^{k_{h,n}(s,a)}_{h+1})(s,a) \le L^2_{h+1}.$$

Moreover, for every k, h, s, a, $\sum_{n=1}^{N_h^k(s,a)} \alpha_n^{N_h^k(s,a)} (\mathbb{P}_h e_{h+1}^{k_{h,n}(s,a)})(s,a)$ is $\mathcal{D}_{<k}$-measurable because $k_{h,n}(s,a) < k$ for every $1 \le n \le N_h^k(s,a)$.

By Lemma D.1, $0 \le V_{h+1}^k \le L_{h+1}$. Moreover, V_{h+1}^k is $\mathcal{D}_{<k}$-measurable, since the iterate before episode k depends only on $\tau^1, \dots, \tau^{k-1}$. Therefore, Lemma B.3 also holds for KLPQ. Apply it with failure probability $\delta/4$ and denote the resulting event by $\mathcal{E}_1$. Choose the fixed universal constant c_b so that $c_b \ge 4 \max\{c, 1\}$, where c is the universal constant in Lemma B.3. For $N \ge 1$, Lemma B.1 with exponent $1/2$ gives

$$c_b L_{h+1} \iota \sqrt{\frac{H}{N}} \le \sum_{n=1}^{N} \alpha_n^N b_{h,n} \le 2 c_b L_{h+1} \iota \sqrt{\frac{H}{N}}. \tag{31}$$

Moreover, on $\mathcal{E}_1$, for $1 \le N \le N_h^{K+1}(s,a)$,

$$\left| \sum_{n=1}^{N} \alpha_n^N (\widehat{\mathbb{P}}_h^{k_{h,n}(s,a)} - \mathbb{P}_{h,s,a}) V_{h+1}^{k_{h,n}(s,a)} \right| \le 4 c L_{h+1} \iota \sqrt{\frac{H}{N}} \le c_b L_{h+1} \iota \sqrt{\frac{H}{N}} \le \sum_{n=1}^{N} \alpha_n^N b_{h,n}. \tag{32}$$

The first inequality is immediate when $L_{h+1} = 0$. Otherwise, $H \ge 2$, so $\iota \ge \log 2$ and $\log(4SAT/\delta) = \iota + \log 4 \le 4\iota$. The second follows from $c_b \ge 4c$, and the last is the lower bound in (31). For $N = 0$, both weighted sums are zero.

Lemma D.2. *On the event $\mathcal{E}_1$, if c_b is fixed as above,*

$$\widetilde{Q}_h^j(s,a) \le Q_h^k(s,a) \le Q_h^{\pi^k}(s,a) \le Q_h^\star(s,a), \qquad V_h^k(s) \le V_h^{\pi^k}(s) \le V_h^\star(s) \tag{33}$$

simultaneously for every $k \in [K+1]$, $j \in [k]$, $h \in [H]$, and $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Proof. We work on the event $\mathcal{E}_1$ defined above. Fix $k \in [K+1]$ and proceed by backward induction on h. For the base case $h = H+1$, $V_{H+1}^{\pi^k} = V_{H+1}^k = V_{H+1}^\star = 0$.

For the induction hypothesis, suppose that the desired value ordering holds at $h+1$, namely, $V_{h+1}^k \le V_{h+1}^{\pi^k} \le V_{h+1}^\star$. For the induction step from $h+1$ to h, the induction hypothesis, monotonicity, and nonnegativity give $V_{h+1}^{\pi^k} \ge V_{h+1}^k \ge V_{h+1}^i \ge 0$ for every $i \le k$, and therefore $Q_h^{\pi^k}(s,a) \ge 0$.

Fix $j \le k$ and (s,a), write $N := N_h^j(s,a)$, and suppress the dependence of $k_{h,n}$ on (s,a). Unrolling Line 11 gives

$$\widetilde{Q}_h^j(s,a) = \alpha_0^N \widetilde{Q}_h^1(s,a) + \sum_{n=1}^{N} \alpha_n^N \left(r_h(s,a) + \widehat{\mathbb{P}}_h^{k_{h,n}} V_{h+1}^{k_{h,n}} - b_{h,n} \right). \tag{34}$$

Writing the Bellman equation for $Q_h^{\pi^k}(s,a)$ with the same weights and subtracting (34) yields

$$\begin{aligned} Q_h^{\pi^k}(s,a) - \widetilde{Q}_h^j(s,a) &= \alpha_0^N Q_h^{\pi^k}(s,a) + \sum_{n=1}^{N} \alpha_n^N \mathbb{P}_{h,s,a}(V_{h+1}^{\pi^k} - V_{h+1}^{k_{h,n}}) \\ &\quad + \sum_{n=1}^{N} \alpha_n^N b_{h,n} - \sum_{n=1}^{N} \alpha_n^N (\widehat{\mathbb{P}}_h^{k_{h,n}} - \mathbb{P}_{h,s,a}) V_{h+1}^{k_{h,n}} \\ &\ge \alpha_0^N Q_h^{\pi^k}(s,a) + \sum_{n=1}^{N} \alpha_n^N \mathbb{P}_{h,s,a}(V_{h+1}^{\pi^k} - V_{h+1}^{k_{h,n}}) \ge 0. \end{aligned}$$

The first inequality follows from (32), the second uses $Q_h^{\pi^k} \ge 0$ and $V_{h+1}^{\pi^k} \ge V_{h+1}^{k_{h,n}}$ for every historical visit $k_{h,n} < j \le k$. This also covers $N = 0$, since every sum $\sum_{n=1}^{N}(\cdots)$ above is then zero. Thus $\widetilde{Q}_h^j(s,a) \le Q_h^{\pi^k}(s,a)$ for every $j \le k$, and hence

$$Q_h^k(s,a) = \max_{1 \le j \le k} \widetilde{Q}_h^j(s,a) \le Q_h^{\pi^k}(s,a) \le Q_h^\star(s,a),$$

where the last inequality follows from Proposition 2.1. Since $\pi_h^k(\cdot \mid s) = G_{h,s}(Q_h^k(s,\cdot))$,

$$\begin{aligned} V_h^k(s) &= \langle \pi_h^k(\cdot \mid s), Q_h^k(s,\cdot)\rangle_{\mathcal{A}} - \frac{1}{\eta}\mathrm{KL}\big(\pi_h^k(\cdot \mid s)\big\|\pi_h^{\mathrm{ref}}(\cdot \mid s)\big) \\ &\le \left\langle \pi_h^k(\cdot \mid s), Q_h^{\pi^k}(s,\cdot)\right\rangle_{\mathcal{A}} - \frac{1}{\eta}\mathrm{KL}\big(\pi_h^k(\cdot \mid s)\big\|\pi_h^{\mathrm{ref}}(\cdot \mid s)\big) \\ &= V_h^{\pi^k}(s) \le V_h^\star(s). \end{aligned}$$

This proves the induction step at stage h and hence, by backward induction, equation (33). Since all arguments beyond the deterministic range and monotonicity properties use only $\mathcal{E}_1$, this completes the proof. □

Lemma D.3. *On $\mathcal{E}_1$, for every $k \in [K+1]$, $h \in [H]$, and $s \in \mathcal{S}$,*

$$(V_h^\star(s) - V_h^k(s))^2 \le e_h^k(s). \tag{35}$$

Moreover, the output policy in Algorithm 2 satisfies

$$\mathrm{SubOpt}(\widehat{\pi}) \le \eta \sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d_h^{\pi^\star}}\left[(Q_h^\star(s,a) - Q_h^{K+1}(s,a))^2\right]. \tag{36}$$

Proof. Work on $\mathcal{E}_1$. For the value error, convexity of $F_{h,s}$ at $Q_h^\star(s,\cdot)$ and the ordering $Q_h^k \le Q_h^\star$ give

$$0 \le V_h^\star(s) - V_h^k(s) \le \langle \pi_h^\star(\cdot \mid s), Q_h^\star(s,\cdot) - Q_h^k(s,\cdot)\rangle_{\mathcal{A}}.$$

Squaring and applying Jensen's inequality yields

$$(V_h^\star(s) - V_h^k(s))^2 \le \sum_a \pi_h^\star(a \mid s)[Q_h^\star(s,a) - Q_h^k(s,a)]^2 = e_h^k(s),$$

which proves the first claim.

It remains to prove the output-policy bound. Let $\widehat{\pi} = \pi^{K+1}$. By (33),

$$0 \le Q_h^{\widehat{\pi}}(s,a) - Q_h^{K+1}(s,a) \le Q_h^\star(s,a) - Q_h^{K+1}(s,a).$$

By Bellman consistency,

$$\begin{aligned} V_h^\star(s) &= \langle \pi_h^\star(\cdot \mid s), Q_h^\star(s,\cdot)\rangle_{\mathcal{A}} - \frac{1}{\eta}\mathrm{KL}\big(\pi_h^\star(\cdot \mid s)\big\|\pi_h^{\mathrm{ref}}(\cdot \mid s)\big), \\ V_h^{\widehat{\pi}}(s) &= \left\langle \widehat{\pi}_h(\cdot \mid s), Q_h^{\widehat{\pi}}(s,\cdot)\right\rangle_{\mathcal{A}} - \frac{1}{\eta}\mathrm{KL}\big(\widehat{\pi}_h(\cdot \mid s)\big\|\pi_h^{\mathrm{ref}}(\cdot \mid s)\big). \end{aligned}$$

Subtracting the two identities and adding and subtracting $\langle \pi_h^\star(\cdot \mid s), Q_h^{\widehat{\pi}}(s,\cdot)\rangle_{\mathcal{A}}$ yields

$$\begin{aligned} V_h^\star(s) - V_h^{\widehat{\pi}}(s) &= \left\langle \pi_h^\star(\cdot \mid s) - \widehat{\pi}_h(\cdot \mid s), Q_h^{\widehat{\pi}}(s,\cdot)\right\rangle_{\mathcal{A}} + \left\langle \pi_h^\star(\cdot \mid s), Q_h^\star(s,\cdot) - Q_h^{\widehat{\pi}}(s,\cdot)\right\rangle_{\mathcal{A}} \\ &\quad - \frac{1}{\eta}\mathrm{KL}\big(\pi_h^\star(\cdot \mid s)\big\|\pi_h^{\mathrm{ref}}(\cdot \mid s)\big) + \frac{1}{\eta}\mathrm{KL}\big(\widehat{\pi}_h(\cdot \mid s)\big\|\pi_h^{\mathrm{ref}}(\cdot \mid s)\big) \\ &= \left\langle \pi_h^\star(\cdot \mid s) - \widehat{\pi}_h(\cdot \mid s), Q_h^{\widehat{\pi}}(s,\cdot)\right\rangle_{\mathcal{A}} + \left\langle \pi_h^\star(\cdot \mid s), \mathbb{P}_h(V_{h+1}^\star - V_{h+1}^{\widehat{\pi}})(s,\cdot)\right\rangle_{\mathcal{A}} \\ &\quad - \frac{1}{\eta}\mathrm{KL}\big(\pi_h^\star(\cdot \mid s)\big\|\pi_h^{\mathrm{ref}}(\cdot \mid s)\big) + \frac{1}{\eta}\mathrm{KL}\big(\widehat{\pi}_h(\cdot \mid s)\big\|\pi_h^{\mathrm{ref}}(\cdot \mid s)\big). \end{aligned}$$

Taking expectation of this identity with respect to $s \sim d_h^{\pi^\star}$ and summing over h, using $V_{H+1}^\star = V_{H+1}^{\widehat{\pi}} = 0$, gives

$$\begin{aligned} \mathrm{SubOpt}(\widehat{\pi}) = \sum_{h=1}^{H} \mathbb{E}_{s\sim d_h^{\pi^\star}}\Bigg[&\left\langle \pi_h^\star(\cdot \mid s) - \widehat{\pi}_h(\cdot \mid s), Q_h^{\widehat{\pi}}(s,\cdot)\right\rangle_{\mathcal{A}} \\ &- \frac{1}{\eta}\mathrm{KL}\big(\pi_h^\star(\cdot \mid s)\big\|\pi_h^{\mathrm{ref}}(\cdot \mid s)\big) + \frac{1}{\eta}\mathrm{KL}\big(\widehat{\pi}_h(\cdot \mid s)\big\|\pi_h^{\mathrm{ref}}(\cdot \mid s)\big)\Bigg]. \end{aligned}$$

The Gibbs formula for $\widehat{\pi}_h$ gives

$$\log \frac{\widehat{\pi}_h(a \mid s)}{\pi_h^{\mathrm{ref}}(a \mid s)} = \eta Q_h^{K+1}(s,a) - \log \sum_b \pi_h^{\mathrm{ref}}(b \mid s) e^{\eta Q_h^{K+1}(s,b)}.$$

Hence, we have

$$\begin{aligned}
\frac{1}{\eta}\mathrm{KL}(\pi_h^\star(\cdot \mid s) \| \widehat{\pi}_h(\cdot \mid s)) &= \frac{1}{\eta}\mathrm{KL}\big(\pi_h^\star(\cdot \mid s) \big\| \pi_h^{\mathrm{ref}}(\cdot \mid s)\big) - \big\langle \pi_h^\star(\cdot \mid s), Q_h^{K+1}(s,\cdot)\big\rangle_{\mathcal{A}} \\
&\quad + \frac{1}{\eta}\log \sum_b \pi_h^{\mathrm{ref}}(b \mid s) e^{\eta Q_h^{K+1}(s,b)} \\
&\overset{(i)}{=} \frac{1}{\eta}\mathrm{KL}\big(\pi_h^\star(\cdot \mid s) \big\| \pi_h^{\mathrm{ref}}(\cdot \mid s)\big) - \frac{1}{\eta}\mathrm{KL}\big(\widehat{\pi}_h(\cdot \mid s) \big\| \pi_h^{\mathrm{ref}}(\cdot \mid s)\big) \\
&\quad - \big\langle \pi_h^\star(\cdot \mid s) - \widehat{\pi}_h(\cdot \mid s), Q_h^{K+1}(s,\cdot)\big\rangle_{\mathcal{A}},
\end{aligned}$$

where (i) uses the Gibbs variational identity. Substituting this identity into the preceding expression for $\mathrm{SubOpt}(\widehat{\pi})$ gives

$$\begin{aligned}
\mathrm{SubOpt}(\widehat{\pi}) &= \sum_{h=1}^H \mathbb{E}_{s\sim d_h^{\pi^\star}}\Bigg[\Big\langle \pi_h^\star(\cdot \mid s) - \widehat{\pi}_h(\cdot \mid s), Q_h^{\widehat{\pi}}(s,\cdot) - Q_h^{K+1}(s,\cdot)\Big\rangle_{\mathcal{A}} \\
&\qquad\qquad - \frac{1}{\eta}\mathrm{KL}(\pi_h^\star(\cdot \mid s) \| \widehat{\pi}_h(\cdot \mid s))\Bigg] \\
&\le \sum_{h=1}^H \mathbb{E}_{s\sim d_h^{\pi^\star}}\Big[\Big\langle \pi_h^\star(\cdot \mid s) - \widehat{\pi}_h(\cdot \mid s), Q_h^{\widehat{\pi}}(s,\cdot) - Q_h^{K+1}(s,\cdot)\Big\rangle_{\mathcal{A}}\Big],
\end{aligned}$$

where the inequality uses the nonnegativity of KL divergence.

For every action with $\pi_h^\star(a \mid s) > 0$, the two Gibbs formulas give

$$\begin{aligned}
\pi_h^\star(a \mid s) - \widehat{\pi}_h(a \mid s) &= \pi_h^\star(a \mid s)\left(1 - \frac{e^{-\eta[Q_h^\star(s,a) - Q_h^{K+1}(s,a)]}}{\sum_b \pi_h^\star(b \mid s) e^{-\eta[Q_h^\star(s,b) - Q_h^{K+1}(s,b)]}}\right) \\
&\overset{(i)}{\le} \pi_h^\star(a \mid s)\left(1 - e^{-\eta[Q_h^\star(s,a) - Q_h^{K+1}(s,a)]}\right) \\
&\overset{(ii)}{\le} \eta \pi_h^\star(a \mid s)[Q_h^\star(s,a) - Q_h^{K+1}(s,a)]. \qquad (37)
\end{aligned}$$

Here (i) uses $Q_h^{K+1} \le Q_h^\star$, which implies $\sum_b \pi_h^\star(b \mid s) e^{-\eta[Q_h^\star(s,b) - Q_h^{K+1}(s,b)]} \le 1$, and (ii) uses $1 - e^{-v} \le v$ for $v \ge 0$. For actions outside the common support, both probabilities are zero, so (37) holds for every action. Since $Q_h^{\widehat{\pi}}(s,a) - Q_h^{K+1}(s,a) \ge 0$, applying (37) yields

$$\begin{aligned}
\mathrm{SubOpt}(\widehat{\pi}) &\le \eta \sum_{h=1}^H \mathbb{E}_{s\sim d_h^{\pi^\star}} \sum_a \pi_h^\star(a \mid s)[Q_h^\star(s,a) - Q_h^{K+1}(s,a)][Q_h^{\widehat{\pi}}(s,a) - Q_h^{K+1}(s,a)] \\
&\le \eta \sum_{h=1}^H \mathbb{E}_{s\sim d_h^{\pi^\star}} \sum_a \pi_h^\star(a \mid s)[Q_h^\star(s,a) - Q_h^{K+1}(s,a)]^2 \\
&= \eta \sum_{h=1}^H \mathbb{E}_{(s,a)\sim d_h^{\pi^\star}}\big[(Q_h^\star(s,a) - Q_h^{K+1}(s,a))^2\big],
\end{aligned}$$

where the second inequality uses $Q_h^{\widehat{\pi}}(s,a) - Q_h^{K+1}(s,a) \le Q_h^\star(s,a) - Q_h^{K+1}(s,a)$ from (33), and the last equality follows from $d_h^{\pi^\star}(s,a) = d_h^{\pi^\star}(s)\pi_h^\star(a \mid s)$. This proves the second claim. □

Lemma D.4. *On $\mathcal{E}_1$, simultaneously for every $k \in [K+1]$, $h \in [H]$, and $(s,a) \in \mathcal{S} \times \mathcal{A}$, writing $N := N_h^k(s,a)$,*

$$\begin{aligned}
(\Delta_h^k(s,a))^2 &\le L_h^2 \mathbf{1}\{N = 0\} + cH^2 L_{h+1}^2 \iota^2 \frac{\mathbf{1}\{N \ge 1\}}{N \vee 1} \\
&\quad + \left(1 + \frac{1}{H}\right) \sum_{n=1}^N \alpha_n^N (\mathbb{P}_h e_{h+1}^{k_{h,n}(s,a)})(s,a),
\end{aligned}$$

where $c > 0$ is a universal constant.

Proof. Fix k, h, s, a, let $N := N_h^k(s,a)$, and suppress the dependence of $k_{h,n}$ on (s,a). If $N = 0$, then $Q_h^k(s,a) = 0$ and $(\Delta_h^k(s,a))^2 \le L_h^2$ by (1).

Suppose $N \ge 1$. Subtracting (34) with $j = k$ from the optimal Bellman equation gives

$$Q_h^\star(s,a) - \widetilde{Q}_h^k(s,a) = \sum_{n=1}^{N} \alpha_n^N \mathbb{P}_{h,s,a}(V_{h+1}^\star - V_{h+1}^{k_{h,n}}) + \sum_{n=1}^{N} \alpha_n^N b_{h,n} - \sum_{n=1}^{N} \alpha_n^N (\widehat{\mathbb{P}}_h^{k_{h,n}} - \mathbb{P}_{h,s,a}) V_{h+1}^{k_{h,n}},$$

since $\alpha_0^N = 0$. By (33) and (32),

$$0 \le Q_h^\star(s,a) - Q_h^k(s,a) \le Q_h^\star(s,a) - \widetilde{Q}_h^k(s,a) \le \sum_{n=1}^{N} \alpha_n^N \mathbb{P}_{h,s,a}(V_{h+1}^\star - V_{h+1}^{k_{h,n}}) + 2\sum_{n=1}^{N} \alpha_n^N b_{h,n}.$$

Consequently,

$$\begin{aligned}(\Delta_h^k(s,a))^2 &\le \left(1 + \frac{1}{H}\right)\left(\sum_{n=1}^{N} \alpha_n^N \mathbb{P}_{h,s,a}(V_{h+1}^\star - V_{h+1}^{k_{h,n}})\right)^2 + 4(H+1)\left(\sum_{n=1}^{N} \alpha_n^N b_{h,n}\right)^2 \\ &\le \left(1 + \frac{1}{H}\right)\sum_{n=1}^{N} \alpha_n^N \mathbb{P}_{h,s,a}\left[(V_{h+1}^\star - V_{h+1}^{k_{h,n}})^2\right] + \frac{32 c_b^2 H^2 L_{h+1}^2 \iota^2}{N} \\ &\le \left(1 + \frac{1}{H}\right)\sum_{n=1}^{N} \alpha_n^N (\mathbb{P}_h e_{h+1}^{k_{h,n}})(s,a) + \frac{c H^2 L_{h+1}^2 \iota^2}{N}.\end{aligned}$$

The first inequality uses $(x+y)^2 \le (1+1/H)x^2 + (H+1)y^2$. The second follows from Jensen's inequality, first with respect to the weights $\{\alpha_n^N\}$ and then with respect to $\mathbb{P}_h(\cdot \mid s,a)$, together with the upper bound in (31). The last inequality follows from (35), after enlarging the universal constant c to dominate $32c_b^2$. Together with the zero-count case, this proves the lemma. □

We now introduce common event for the proof of Theorem 4.1: If $SAT > 1$, apply Lemma B.4 with failure probability $\delta/4$ and denote its event by $\mathcal{E}_2$, if $SAT = 1$, let $\mathcal{E}_2$ be the sure event. Apply Lemma B.5 twice with failure probability $\delta/4$, using

$$U_h^k(s,a) = \sum_{n=1}^{N_h^k(s,a)} \alpha_n^{N_h^k(s,a)} (\mathbb{P}_h e_{h+1}^{k_{h,n}(s,a)})(s,a) \quad \text{and} \quad U_h^k(s,a) = (\mathbb{P}_h e_{h+1}^k)(s,a),$$

respectively, with e_h^k defined as in Section 5.1 for Algorithm 2. The deterministic bounds and predictability above ensure that both choices of $U_h^k(s,a)$ with $B_h = L_{h+1}^2$ satisfy the conditions of Lemma B.5. Denote the two events by $\mathcal{E}_3$ and $\mathcal{E}_4$, and set

$$\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4, \qquad \Pr(\mathcal{E}) \ge 1 - \delta.$$

All four events are defined for the same iterates generated using the original bonus in (2).

D.2 Proof of Theorem 4.1

Proof. If $SAT = 1$, the output is the unique policy, so $\mathrm{SubOpt}(\widehat{\pi}) = 0$. Hence suppose $SAT > 1$. Work on the common event $\mathcal{E}$ defined above, which has probability at least $1 - \delta$. Starting from Lemma D.4, we repeat the argument used in the proof of Lemma C.4. The count-control step, the two applications of Lemma B.5, and the reindexing argument are unchanged, up to universal constants, because the required boundedness and predictability conditions are verified in Appendix D.1.

The only difference is that the propagated error term in Lemma D.4 has no factor γ_{h+1}. Therefore, the same calculation yields, for every $h \in [H]$,

$$\sum_{k=1}^{K} \sum_{s} d_h^{\pi^\star}(s) e_h^k(s) \le \left(1 + \frac{5}{H}\right) \sum_{k=1}^{K} \sum_{s} d_{h+1}^{\pi^\star}(s) e_{h+1}^k(s) + c_0 SAC_\eta^\star H^2 L_h^2 \iota^3 (1 + \log K).$$

Applying the same backward iteration argument as in the proof of Lemma C.4, now without the products of γ_j, and using $e_{H+1}^k \equiv 0$ together with

$$\left(1 + \frac{5}{H}\right)^H \le e^5, \qquad \sum_{t=h}^{H} L_t^2 \le L_h^3,$$

gives

$$\sum_{k=1}^{K} \mathbb{E}_{(s,a)\sim d_h^{\pi^\star}} \left[(Q_h^k(s,a) - Q_h^\star(s,a))^2 \right] \le c_1 SAC_\eta^\star H^2 L_h^3 \iota^3 (1 + \log K),$$

which gives the desired accumulated squared Q-estimation error bound.

By Line 12, $Q_h^{k+1} \ge Q_h^k$ coordinatewise, while equation (33) gives $Q_h^k \le Q_h^\star$. Therefore, for every $k \in [K]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$0 \le Q_h^\star(s,a) - Q_h^{K+1}(s,a) \le Q_h^\star(s,a) - Q_h^k(s,a).$$

Squaring, averaging first with respect to $d_h^{\pi^\star}$, and then averaging over $k \in [K]$ gives

$$\begin{aligned} \mathbb{E}_{(s,a)\sim d_h^{\pi^\star}} \left[(Q_h^{K+1}(s,a) - Q_h^\star(s,a))^2 \right] &\le \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{(s,a)\sim d_h^{\pi^\star}} \left[(Q_h^k(s,a) - Q_h^\star(s,a))^2 \right] \\ &\le c_1 \frac{SAC_\eta^\star}{K} H^2 L_h^3 \iota^3 (1 + \log K). \end{aligned}$$

Substituting this into (36) gives

$$\begin{aligned} \mathrm{SubOpt}(\widehat{\pi}) &\le c \frac{\eta SAC_\eta^\star H^2 \iota^3 (1 + \log K)}{K} \sum_{h=1}^{H} L_h^3 \\ &\le c \frac{\eta H^6 SAC_\eta^\star}{K} \iota^3 (1 + \log K). \end{aligned}$$

The second inequality uses $\sum_{h=1}^{H} L_h^3 \le H^4$. This proves Theorem 4.1. □