Online KL-Regularized Reinforcement Learning with Function Approximation under Misspecification

Haoyang Hong*, Zichen Wang*, Quanquan Gu, Huazheng Wang

Keywords: KL regularization, contextual bandits, reinforcement learning, model misspecification

Summary

We study KL-regularized contextual bandits and episodic reinforcement learning with reference policies under general function approximation and model misspecification, motivated by KL-constrained policy optimization and RLHF-style post-training (Schulman et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023; Zhao et al., 2025). The key difficulty is that KL regularization changes the benchmark, so standard misspecification conditions do not transfer directly. We introduce KL misspecification conditions and analyze optimistic regression-based algorithms with Gibbs policy updates. Our main results are a high-probability KL-regret guarantee for contextual bandits and a modular leading-order high-probability regret theorem for episodic RL under explicit confidence and uncertainty conditions, both with explicit dependence on misspecification and statistical complexity, recovering the corresponding realizable guarantees of Zhao et al. (2025) as special cases.

Contribution(s)

1. We introduce KL-aligned misspecification conditions for KL-regularized contextual bandits and episodic reinforcement learning under general function approximation.
   Context: The conditions are motivated by the Gibbs form of the KL-regularized objective and complement prior KL-regularized analyses that focus on realizable or near-realizable settings (Zhao et al., 2025). Our contextual-bandit condition is a pointwise misspecification control over all state-action pairs, paralleling classical contextual-bandit misspecification (Foster et al., 2021b; Takemura et al., 2021). Our RL condition extends this pointwise control to stagewise KL Bellman-backup targets.
2. We study oracle-based optimistic algorithms for KL-regularized contextual bandits and episodic reinforcement learning, combining regression-based estimation, uncertainty bonuses, and Gibbs policy updates relative to a reference policy.
   Context: The algorithms follow the optimistic reward-estimation template and are compatible with the KL-regularized objective. Our primary guarantees are for the known-misspecification setting, where the misspecification levels enter the bonuses as inputs.
3. We establish high-probability KL-regret guarantees with explicit misspecification terms and complexity-sensitive dependence, including a direct bandit theorem and a high-probability regret theorem for episodic RL stated under assumed confidence and uncertainty conditions.
   Context: The guarantees are stated against the KL-optimal Gibbs comparator and use localized eluder-dimension-style complexity terms. For RL, the theorem is modular and is under explicit confidence and uncertainty conditions.

Online KL-Regularized Reinforcement Learning with Function Approximation under Misspecification

Haoyang Hong[1]*, **Zichen Wang**[2]*, **Quanquan Gu**[3], **Huazheng Wang**[1]
honghao@oregonstate.edu, zichenw6@illinois.edu, qgu@cs.ucla.edu, huazheng.wang@oregonstate.edu

[1]School of Electrical Engineering and Computer Science, Oregon State University
[2]Department of Electrical and Computer Engineering and Coordinated Science Laboratory, University of Illinois Urbana-Champaign
[3]Department of Computer Science, University of California, Los Angeles

Abstract

We study KL-regularized contextual bandits and episodic reinforcement learning (RL) under general function approximation with model misspecification. Existing guarantees rely on realizability and therefore do not extend to misspecified models, where classical regret bounds may fail. This work introduces KL misspecification formulations for contextual bandits and episodic RL and analyzes regression-based algorithms with Gibbs policy updates. High-probability KL-regret guarantees with explicit misspecification terms are established, recovering the standard realizable KL-regularized setting as a special case.

1 Introduction

We study KL-regularized contextual bandits and episodic RL under general function approximation with model misspecification. Such formulations arise in modern RLHF and LLM post-training pipelines, where policy updates are regularized by a KL penalty relative to a reference policy (Schulman et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023; Xiong et al., 2024; Zhao et al., 2024; 2025). Beyond their practical relevance, KL-regularized objectives provide a principled abstraction of stability–performance trade-offs induced by information-theoretic regularization.

Informally, model misspecification means that the learner uses a model class that is too simple to exactly represent the environment. For example, the learner may fit rewards or value targets using a linear, low-rank, or otherwise restricted class, while the true reward depends on nonlinear features or interactions outside that class. In that case, even with unlimited data and no statistical noise, the best predictor inside the chosen class can still have a nonzero approximation bias. This is the failure of realizability studied in misspecified bandits and RL (Foster et al., 2021b; Krishnamurthy et al., 2021; Takemura et al., 2021; Vial et al., 2022; Li & Yang, 2024).

Our analysis builds on the literature on contextual bandits and RL with general function approximation, which develops oracle-based and regression-based algorithms together with strong regret guarantees under structural assumptions (Russo & Van Roy, 2013; Agarwal et al., 2014; Foster et al., 2018; Foster & Rakhlin, 2020; Wang et al., 2020; Jin et al., 2021; Foster et al., 2021a; Xie et al., 2023). Subsequent work has further expanded this framework to include policy-optimization-oriented analyses and more general oracle-efficient formulations (Levy & Mansour, 2026; Levy et al., 2025; Qin & Zhang, 2026). A separate recent line extends these ideas to KL-regularized bandit and RL settings, establishing provably efficient guarantees for soft-policy formulations (Xiong

*Equal contribution.

et al., 2024; Xie et al., 2024; Cen et al., 2024; Zhao et al., 2024; 2025; Wu et al., 2025; Lee et al., 2026). However, these analyses rely on realizability, requiring the ground-truth reward or value functions to lie within the chosen function class. When this assumption fails, the resulting misspecified setting can fundamentally alter achievable regret guarantees.

Recent work on misspecified bandits and RL shows that achievable guarantees depend critically on the chosen misspecification model (Foster et al., 2021b; Krishnamurthy et al., 2021; Takemura et al., 2021; Vial et al., 2022; Li & Yang, 2024; Amortila et al., 2024). In KL-regularized problems, this dependence is further complicated by structural differences from standard reward-regret formulations. First, the relevant performance criterion is KL regret with respect to a KL-optimal Gibbs policy, rather than reward regret relative to a deterministic benchmark. Second, the KL regularization changes both the optimization geometry and the Bellman targets appearing in the analysis, so misspecification conditions from standard bandit/RL formulations do not transfer directly. Existing misspecification models do not explicitly account for these KL-specific features, leaving the theoretical treatment of misspecification in KL-regularized bandits and RL incomplete.

This raises the question of whether one can obtain provably efficient KL-regret guarantees under misspecification using approximation conditions aligned with the KL-regularized objective. We address this question for KL-regularized contextual bandits and episodic RL. For RL, our main theorem is stated under explicitly assumed confidence and uncertainty conditions, and the algorithmic guarantees use bonuses calibrated to a known misspecification level. The main contributions are as follows:

- We introduce KL misspecification formulations for contextual bandits and episodic RL. For contextual bandits, we use a pointwise misspecification formulation adapted to the KL-regularized objective. For episodic RL, we introduce a stagewise misspecification condition aligned with KL-regularized Bellman targets.
- The main technical contribution is an analysis that combines a Gibbs quadratic self-bounding inequality for KL gaps with a reduction that converts squared on-policy Q-gaps into squared Bellman residual terms. This isolates the KL-specific terms that are not present in standard misspecified reward-regret analyses.
- We establish high-probability KL-regret guarantees with explicit misspecification terms and eluder-dimension-style complexity dependence. For contextual bandits, we give a direct regret theorem under the KL pointwise misspecification model. For episodic RL, we prove a high-probability regret theorem under assumed confidence and uncertainty conditions, with explicit dependence on misspecification and a bound on the sum of squared bonuses. We further show how our framework recovers standard realizable KL-regularized settings as special cases.

2 Related Work

General function approximation in contextual bandits and RL. Our work builds on the literature on contextual bandits and RL with general function approximation, including reduction-based, oracle-based, and regression-based approaches (Langford & Zhang, 2008; Agarwal et al., 2014; Foster et al., 2018; Foster & Rakhlin, 2020). This literature also develops structural complexity measures and complexity-sensitive analyses that are central to modern online learning theory over rich hypothesis classes (Russo & Van Roy, 2013; Wang et al., 2020; Jin et al., 2021; Foster et al., 2021a). Related work further studies how structural conditions, such as coverage, affect exploration efficiency in online RL (Xie et al., 2023), and recent works continue to extend oracle-efficient and policy-optimization-oriented analyses for contextual bandits with rich function classes (Levy & Mansour, 2026; Levy et al., 2025; Qin & Zhang, 2026).

KL-regularized bandits, RL, and RLHF-related theory. A line of work studies KL-regularized, or relative-entropy-regularized, bandit and RL formulations. Related theoretical developments in preference-based RL and RLHF include finite-time analyses, trajectory-preference formulations,

randomized exploration schemes, and comparisons between RLHF and standard RL (Xu et al., 2020; Pacchiano et al., 2023; Wu & Sun, 2023; Wang et al., 2023). More recent results analyze KL-regularized RLHF formulations directly, including iterative preference learning under KL constraints, exploration-aware preference optimization, online and offline settings, and sharp guarantees for KL-regularized contextual bandits and RL (Xiong et al., 2024; Xie et al., 2024; Cen et al., 2024; Zhao et al., 2024; 2025). Additional work studies general-preference and Nash-style RLHF as well as regularized variants beyond the standard Bradley–Terry model (Munos et al., 2023; Ye et al., 2024; Wu et al., 2025; Lee et al., 2026).

Misspecification in bandits and RL. A substantial body of work studies misspecified contextual bandits and RL, establishing forms of graceful degradation beyond realizability and highlighting the sensitivity of guarantees to the chosen misspecification model (Foster et al., 2021b; Krishnamurthy et al., 2021; Takemura et al., 2021; Vial et al., 2022; Li & Yang, 2024). These developments include oracle-efficient contextual bandit methods, misspecified linear contextual bandits, and misspecified RL analyses spanning linear MDPs to more general function approximation settings (Krishnamurthy et al., 2021; Takemura et al., 2021; Vial et al., 2022; Li & Yang, 2024). Our bandit misspecification condition adopts a pointwise approximation viewpoint extending to the ideas of the misspecification formulations used by Foster et al. (2021b), but adapts the formulation to KL-regularized objectives and extends the pointwise viewpoint to KL-regularized Bellman targets in episodic RL. Related work on misspecified regression under covariate shift provides complementary analytical tools for learning with approximation error and distribution shift (Amortila et al., 2024).

3 Preliminaries

3.1 KL-Regularized Contextual Bandits

Suppose there are a total of T rounds. At each round $t \in [T]$, a context $x_t \in \mathcal{X}$ is drawn from an unknown distribution d, where $\mathcal{X}$ denotes the context space. At round t, the learning algorithm observes x_t and selects a policy $\pi_t : \mathcal{X} \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ denotes the set of distributions over $\mathcal{A}$. An action is sampled according to $a_t \sim \pi_t(\cdot \mid x_t)$. The learner subsequently receives a reward

$$r_t = R^\star(x_t, a_t) + \epsilon_t, \tag{1}$$

where $R^\star : \mathcal{X} \times \mathcal{A} \to [0,1]$ is an unknown ground-truth reward function, and ϵ_t is conditionally zero-mean and 1-sub-Gaussian. We assume the learner is given a reference policy $\pi_{\mathrm{ref}}(\cdot \mid x)$ and a finite function class $\mathcal{R} \subseteq [0,1]^{\mathcal{X}\times\mathcal{A}}$ with cardinality $N_{\mathcal{R}} := |\mathcal{R}|$.

Learning objective We aim to design a learning algorithm that minimizes the regret, which is defined as follows:

$$\mathrm{Reg}(T) = \sum_{t\in[T]} \big(J(\pi^*) - J(\pi_t)\big), \tag{2}$$

where $J(\pi)$ denotes the KL-regularized objective used here (as in Zhang (2023)):

$$\begin{aligned} J(\pi) &:= \mathbb{E}_{x\sim d}\, \mathbb{E}_{a\sim\pi(\cdot|x)} \left[R^\star(x,a) - \eta^{-1}\mathrm{KL}(\pi(\cdot \mid x) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid x))\right] \\ &= \mathbb{E}_{x\sim d}\, \mathbb{E}_{a\sim\pi(\cdot|x)} \left[R^\star(x,a) - \eta^{-1} \log \frac{\pi(a \mid x)}{\pi_{\mathrm{ref}}(a \mid x)}\right], \end{aligned} \tag{3}$$

where $\eta > 0$ is the regularization parameter. Here $\pi(a \mid x) : \mathcal{X}\times\mathcal{A} \to [0,1]$ denotes the conditional probability of selecting a given context x. We define the optimal policy $\pi^* = \arg\max_\pi J(\pi)$, and π_t to denote the stochastic policy adopted by the learning algorithm at round t.

Throughout this subsection, policy optimization is taken over all stochastic policies π. We use the standard convention that $\mathrm{KL}(\pi(\cdot \mid x) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid x)) = +\infty$ whenever $\pi(\cdot \mid x)$ is not absolutely continuous with respect to $\pi_{\mathrm{ref}}(\cdot \mid x)$. Thus, only policies that are absolutely continuous with respect to $\pi_{\mathrm{ref}}(\cdot \mid x)$ have finite KL-regularized objective values.

Lemma 3.1 (Solution to the KL-regularized bandits optimization (Zhang, 2023))**.** *For any $x \in \mathcal{X}$ and any measurable function $R : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ such that $\exp(\eta R(x, \cdot))$ is integrable under $\pi_{\mathrm{ref}}(\cdot \mid x)$, we have*

$$\begin{aligned} &\max_{\pi} \mathbb{E}_{a\sim\pi(\cdot|x)}\left[R(x,a) - \eta^{-1}\mathrm{KL}(\pi(\cdot \mid x) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid x))\right] \\ &\quad = \eta^{-1} \log \mathbb{E}_{a\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\eta R(x,a))]. \end{aligned} \tag{4}$$

The maximizer is given by

$$\pi_R(a \mid x) = \frac{1}{Z_R(x)} \pi_{\mathrm{ref}}(a \mid x) \exp\left(\eta R(x,a)\right), \tag{5}$$

where $Z_R(x) := \mathbb{E}_{a'\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\eta R(x,a'))]$.

Using Lemma 3.1 and (3), the KL-optimal policy is $\pi^* = \pi_{R^\star}$. Importantly, this characterization does not require realizability; in our misspecified setting, $\mathcal{R}$ only specifies the learner's approximation class, while the benchmark policy is defined by the ground-truth reward $R^\star$.

Assumption 3.2 (Misspecification for contextual bandits)**.** *Define the pointwise misspecification for the bandit reward by*

$$\zeta^2_{\mathrm{Bandit}} := \inf_{R\in\mathcal{R}} \sup_{x\in\mathcal{X}} \sup_{a\in\mathcal{A}} \big(R(x,a) - R^\star(x,a)\big)^2. \tag{6}$$

This is the KL-regularized extension of the pointwise misspecification formulation in Foster et al. (2021b): the benchmark and regret notion are KL-regularized, while the approximation error is quantified in the same pointwise style.

3.2 KL-Regularized RL

In this section, we introduce the episodic KL-regularized MDP formulation. An episodic MDP is defined by a tuple $(\mathcal{S}, \mathcal{A}, H, T, \mathbb{P}, d, r)$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ denotes the action space, H denotes the episode length, and T denotes the total number of episodes. The transition kernel is given by $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$, where $\mathbb{P}_h(s_{h+1} \mid s_h, a_h)$ denotes the probability of transitioning from the current state-action pair (s_h, a_h) to the next state s_{h+1} at step h. The initial state s_1^t is drawn from an unknown distribution d, and the reward function is defined as $r = \{r_h : \mathcal{S} \times \mathcal{A} \to [0,1]\}_{h=1}^H$. For each round $t \in [T]$ and each step $h \in [H]$, the learning algorithm observes the current state s_h^t, takes action a_h^t, receives the reward $r_h^t(s_h^t, a_h^t)$, and transitions to the next state s_{h+1}^t according to the transition kernel $\mathbb{P}_h(\cdot \mid s_h^t, a_h^t)$. In the value-function definitions and Bellman-operator definitions below, $r_h(s,a)$ denotes the underlying stage-h reward function in the MDP model, while $r_h^t(s_h^t, a_h^t)$ denotes the reward observed in episode t.

We use stagewise reference policies $\{\pi_{\mathrm{ref},h}\}_{h=1}^H$. As in the bandit case, policy optimization is taken over all stochastic policies, with the standard convention that $\mathrm{KL}(\pi_h(\cdot \mid s) \,\|\, \pi_{\mathrm{ref},h}(\cdot \mid s)) = +\infty$ whenever $\pi_h(\cdot \mid s)$ is not absolutely continuous with respect to $\pi_{\mathrm{ref},h}(\cdot \mid s)$. Thus, only policies that are absolutely continuous with respect to the stagewise reference policies have finite KL-regularized values.

We define the value function and Q-function as follows:

$$\begin{aligned} V_h^\pi(s_h) &= \sum_{j=h}^H \mathbb{E}^\pi \left[r_j(s_j,a_j) - \eta^{-1}\mathrm{KL}(\pi_j(\cdot \mid s_j) \,\|\, \pi_{\mathrm{ref},j}(\cdot \mid s_j)) \,\middle|\, s_h \right], \\ Q_h^\pi(s_h,a_h) &= r_h(s_h,a_h) + \sum_{j=h+1}^H \mathbb{E}^\pi \left[r_j(s_j,a_j) - \eta^{-1}\mathrm{KL}(\pi_j(\cdot \mid s_j) \,\|\, \pi_{\mathrm{ref},j}(\cdot \mid s_j)) \,\middle|\, s_h, a_h \right]. \end{aligned} \tag{7}$$

Here $\pi := \{\pi_h\}_{h=1}^H$ denotes the policy sequence, and $\mathbb{E}^\pi$ denotes the expectation over the trajectory induced by π. We can also define the value function and Q-function recursively as follows. We set the terminal value as $V_{H+1}^\pi(s_{H+1}) = 0$, and for each step $h \in [H]$, we define

$$V_h^\pi(s_h) = \mathbb{E}_{a_h \sim \pi_h(\cdot|s_h)} \left[Q_h^\pi(s_h, a_h) - \eta^{-1} \mathrm{KL}(\pi_h(\cdot \mid s_h) \,\|\, \pi_{\mathrm{ref},h}(\cdot \mid s_h)) \right],$$
$$Q_h^\pi(s_h, a_h) = r_h(s_h, a_h) + \mathbb{E}_{s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, a_h)} \left[V_{h+1}^\pi(s_{h+1}) \right].$$

The optimal value function and Q-function are defined as

$$V_h^*(s_h) = \max_\pi V_h^\pi(s_h), \qquad Q_h^*(s_h, a_h) = \max_\pi Q_h^\pi(s_h, a_h). \tag{8}$$

Assume the optimal policy is achieved at π^*. Using Lemma 3.1 and a backward iteration starting from $V_{H+1}^*(s_{H+1}) = 0$, we have the following proposition.

Proposition 3.3 (Solution to the KL-regularized RL optimization (Zhang, 2023))**.** *The optimal policy is given by*

$$\pi_h^*(a_h \mid s_h) = \frac{1}{Z_h^*(s_h)} \pi_{\mathrm{ref},h}(a_h \mid s_h) \exp(\eta Q_h^*(s_h, a_h)), \tag{9}$$

where $Z_h^*(s_h) := \mathbb{E}_{a \sim \pi_{\mathrm{ref},h}(\cdot|s_h)}[\exp(\eta Q_h^*(s_h, a))]$*. Moreover,*

$$V_h^*(s_h) = \eta^{-1} \log \mathbb{E}_{a \sim \pi_{\mathrm{ref},h}(\cdot|s_h)} \left[\exp(\eta Q_h^*(s_h, a)) \right], \tag{10}$$

and

$$Q_h^*(s_h, a_h) = r_h(s_h, a_h) + \mathbb{E}_{s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, a_h)} \left[V_{h+1}^*(s_{h+1}) \right]. \tag{11}$$

Let $B_h := H - h + 1$*. Under the normalization* $r_h(s, a) \in [0, 1]$ *for all* (h, s, a)*, we have* $Q_h^*(s_h, a_h) \in [0, B_h]$ *for any* $(s_h, a_h) \in \mathcal{S} \times \mathcal{A}$*, and consequently* $V_h^*(s_h) \in [0, B_h]$ *for any* $s_h \in \mathcal{S}$*.*

The Gibbs policy characterization and the log-partition recursion above are the KL-regularized dynamic-programming counterparts of Lemma 3.1, and are standard in KL-regularized RL formulations. We use this reference-policy form because it makes the KL-regularized geometry explicit and aligns with the Gibbs-policy benchmark used in our analysis.

We let $\mathcal{R} := \{\mathcal{R}_h\}_{h=1}^H$ denote the finite stagewise function classes used by the algorithm, where $\mathcal{R}_h \subseteq [0, B_h]^{\mathcal{S} \times \mathcal{A}}$, and denote their cardinalities by $N_{\mathcal{R}_h} := |\mathcal{R}_h| < \infty$.

For any bounded measurable continuation reward function $f_{h+1} : \mathcal{S} \times \mathcal{A} \to [0, B_{h+1}]$, define the stagewise log-partition operator and the KL-regularized Bellman operator by

$$V_{h+1}(f_{h+1}; s_{h+1}) := \eta^{-1} \log \mathbb{E}_{a \sim \pi_{\mathrm{ref},h+1}(\cdot|s_{h+1})} \left[\exp(\eta f_{h+1}(s_{h+1}, a)) \right]. \tag{12}$$

$$(\mathcal{T}_{\eta,h} f_{h+1})(s_h, a_h) := r_h(s_h, a_h) + \mathbb{E}_{s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, a_h)} \left[V_{h+1}(f_{h+1}; s_{h+1}) \right]. \tag{13}$$

Learning objective We aim to design a learning algorithm that minimizes the regret. The objective function is defined as

$$J(\pi) = \mathbb{E}_{s_1 \sim d} \left[V_1^\pi(s_1) \right], \tag{14}$$

and the regret is defined by

$$\mathrm{Reg}(T) = \sum_{t \in [T]} \big(J(\pi^*) - J(\pi_t) \big). \tag{15}$$

Using Proposition 3.3, the KL-optimal policy is the Gibbs policy characterized by $\{Q_h^*\}_{h=1}^H$. We next introduce a stagewise misspecification condition stated along a possibly data-dependent KL Bellman-backup path.

To state the misspecification condition in a pathwise form, let $\widetilde{Q}_{t,h+1}$ denote the stage-$(h+1)$ continuation Q-function used in the episode-t, stage-h KL Bellman backup, with the terminal convention $\widetilde{Q}_{t,H+1} \equiv 0$.

Assumption 3.4 (KL misspecification for RL)**.** *We assume there exists* $\zeta_{\mathrm{RL}} \geq 0$ *such that for every* $t \in [T]$ *and every* $h \in [H]$*,*

$$\inf_{Q \in \mathcal{R}_h} \sup_{s \in \mathcal{S}} \sup_{a \in \mathcal{A}} \left| Q(s,a) - (\mathcal{T}_{\eta,h} \widetilde{Q}_{t,h+1})(s,a) \right| \;\leq\; \zeta_{\mathrm{RL}}. \tag{16}$$

This is a KL-regularized, stagewise pointwise misspecification condition along a Bellman-backup path. It extends the pointwise misspecification viewpoint of Foster et al. (2021b) from bandits to KL-regularized Bellman targets, while being weaker than requiring pointwise approximation uniformly for all continuation functions. Compared with locally bounded misspecification formulations in RL that control approximation error on policy-relevant state-action distributions (e.g., Li & Yang (2024)), (16) is a stagewise pointwise condition along a KL Bellman-backup path, and it plays the same structural role in RL as (6) does in contextual bandits.

4 Algorithms

Our algorithms follow the standard regression-based optimistic template for value-based methods under general function approximation, and are compatible with recent KL-regularized analyses at the level of the Gibbs policy-improvement step. The main difference here is that we allow model misspecification and track it explicitly through additive misspecification terms in the optimism bonuses.

Throughout this section, we work in an oracle-based computational model with ERM oracles and uncertainty-bonus routines. Our guarantees are statistical under this oracle model. We present the known-misspecification versions of the algorithms, where ζ_{Bandit} and ζ_{RL} are treated as known inputs and inserted directly into the bonuses.

To separate regression from uncertainty computation, we maintain two histories: the observed-data history and the state–action history. For bandits,

$$\widetilde{D}_{t-1}^{\mathrm{Bandit}} := \{(x_i, a_i, r_i)\}_{i=1}^{t-1}, \qquad \bar{D}_{t-1}^{\mathrm{Bandit}} := \{(x_i, a_i)\}_{i=1}^{t-1},$$

and for episodic RL, for each $h \in [H]$,

$$\widetilde{D}_{h,t-1}^{\mathrm{RL}} := \{(s_{\tau,h}, a_{\tau,h}, r_{\tau,h}, s_{\tau,h+1})\}_{\tau=1}^{t-1}, \qquad \bar{D}_{h,t-1}^{\mathrm{RL}} := \{(s_{\tau,h}, a_{\tau,h})\}_{\tau=1}^{t-1}.$$

For the RL analysis with recomputed regression labels, we also use a finite continuation-value class $\mathcal{V}_{h+1}$ containing all continuation values $\widetilde{V}_{t,h+1}$ generated by Algorithm 2 at stage $h+1$, with $N_{\mathcal{V}_{h+1}} := |\mathcal{V}_{h+1}|$ and $\mathcal{V}_{H+1} = \{0\}$. This convention is used only to make the frozen-target concentration step uniform over the data-dependent continuation values.

Definition 4.1 (Uncertainty and eluder dimension)**.** *We use eluder-style uncertainty quantities from the general function-approximation literature (Russo & Van Roy (2013); see also Wang et al. (2020); Jin et al. (2021)), with a normalization chosen to match the KL-regularized comparison convention used in Zhao et al. (2025). The definitions below do not require the classes to be finite. We state the main high-probability results in the finite-class case to keep the union bounds transparent; the same proof extends to infinite classes by replacing finite cardinalities with suitable terms.*

Contextual bandits. *Given a context–action history* $\bar{D}_{t-1}^{\mathrm{Bandit}}$*, define*

$$U_{\mathcal{R}}\big(\lambda, x, a; \bar{D}_{t-1}^{\mathrm{Bandit}}\big) := \sup_{R_1, R_2 \in \mathcal{R}} \frac{|R_1(x,a) - R_2(x,a)|}{\sqrt{\lambda + \sum_{i=1}^{t-1} (R_1(x_i,a_i) - R_2(x_i,a_i))^2}}, \tag{17}$$

and

$$d(\mathcal{R}, \lambda, T) := \sup_{x_{1:T},\, a_{1:T}} \sum_{t=1}^{T} \min\Big\{1,\; U_{\mathcal{R}}\big(\lambda, x_t, a_t; \bar{D}_{t-1}^{\mathrm{Bandit}}\big)^2\Big\}. \tag{18}$$

Algorithm 1 MR−KL−UCB

Input: $T, \eta, \beta, \lambda, \pi_{\mathrm{ref}}, \mathcal{R}, \zeta_{\mathrm{Bandit}}$
for $t = 1, \dots, T$ **do**
  Observe context x_t
  Fit $\hat{R}_{t-1} \in \arg\min_{R \in \mathcal{R}} \sum_{i=1}^{t-1} (R(x_i, a_i) - r_i)^2$
  Compute $b_{t-1}(x, a) \leftarrow \min\{1, \beta\, U_{\mathcal{R}}(\lambda, x, a; \bar{D}_{t-1}^{\mathrm{Bandit}}) + \zeta_{\mathrm{Bandit}}\}$
  Set $\widetilde{R}_{t-1}(x, a) \leftarrow \Pi_{[0,1]}\big(\hat{R}_{t-1}(x, a) + b_{t-1}(x, a)\big)$
  Update policy $\pi_t(a \mid x_t) \propto \pi_{\mathrm{ref}}(a \mid x_t) \exp\big(\eta \widetilde{R}_{t-1}(x_t, a)\big)$
  Choose action $a_t \sim \pi_t(\cdot \mid x_t)$ and observe reward r_t
end for

Episodic RL. *For each stage $h \in [H]$, given the stagewise history $\bar{D}_{h,t-1}^{\mathrm{RL}}$, define*

$$U_{\mathcal{R}_h}(\lambda, s, a; \bar{D}_{h,t-1}^{\mathrm{RL}}) := \sup_{R_1, R_2 \in \mathcal{R}_h} \frac{|R_1(s,a) - R_2(s,a)|}{\sqrt{\lambda + \sum_{\tau=1}^{t-1} (R_1(s_{\tau,h}, a_{\tau,h}) - R_2(s_{\tau,h}, a_{\tau,h}))^2}}, \tag{19}$$

the stagewise complexity

$$d(\mathcal{R}_h, \lambda, T) := \sup_{s_{1:T,h},\, a_{1:T,h}} \sum_{t=1}^{T} \min\Big\{1,\ U_{\mathcal{R}_h}(\lambda, s_{t,h}, a_{t,h}; \bar{D}_{h,t-1}^{\mathrm{RL}})^2\Big\}, \tag{20}$$

and the aggregated RL complexity

$$d_{\mathrm{RL}}(\lambda, T) := \sum_{h=1}^{H} d(\mathcal{R}_h, \lambda, T). \tag{21}$$

Algorithms 1 and 2 take (β, λ) as explicit inputs for modularity. In the theorem statements, these parameters are instantiated using localized-eluder calibrations at the usual logarithmic scale, with additional $T\zeta^2$-type misspecification terms.

For convenience, we define clipping as $\Pi_{[0,n]}(x) := \min\{n, \max\{0, x\}\}$.

4.1 Contextual bandits: MR−KL−UCB

We adapt the standard upper confidence bound (UCB) method to the KL-regularized setting and propose Misspecification-Robust KL-UCB (MR−KL−UCB) in Algorithm 1. The algorithm performs optimism at the reward-function level and then samples from the induced Gibbs policy. Concretely, at round t, after observing x_t, the learner fits an ERM predictor on $\widetilde{D}_{t-1}^{\mathrm{Bandit}}$, computes an uncertainty bonus using $U_{\mathcal{R}}$ and $\bar{D}_{t-1}^{\mathrm{Bandit}}$ as in Definition 4.1, forms an optimistic clipped reward estimate, and samples from the Gibbs policy induced by that optimistic estimate. All update formulas are given once in Algorithm 1 and are not repeated in the prose. For the theoretical guarantee in Theorem 5.1, we use $\beta = \Theta\Big(\sqrt{\log(N_{\mathcal{R}} T/\delta)}\Big)$ and defer the exact formulas to (40).

Relative to optimistic contextual bandit methods with greedy action selection, MR−KL−UCB performs optimism at the reward-function level and then samples from the corresponding Gibbs policy.

4.2 Episodic RL: MR−KL−LSVI

We adapt LSVI (Jin et al., 2020) to the KL-regularized setting and propose Misspecification-Robust KL-LSVI (MR−KL−LSVI) in Algorithm 2. The algorithm is a KL-regularized extension of backward fitted-Q planning with optimism. Each episode has two phases: a backward planning pass based on previous-episode data, followed by one rollout of the resulting nonstationary Gibbs policy.

Let $B_h := H - h + 1$. In the backward pass, at each stage h, the learner recomputes regression labels using the current continuation value, fits a stagewise ERM predictor on $\widetilde{D}_{h,t-1}^{\mathrm{RL}}$, computes a

Algorithm 2 MR-KL-LSVI

Input: $T, H, \eta, \beta, \lambda, \{\mathcal{R}_h\}_{h=1}^H, \{\pi_{\mathrm{ref},h}\}_{h=1}^H, \zeta_{\mathrm{RL}}$
for $t = 1, \dots, T$ **do**
  Initialize $\widetilde{V}_{t,H+1}(\cdot) \leftarrow 0$
  for $h = H, H-1, \dots, 1$ **do**
    Set recomputed labels $y_{\tau,h}^{(t)} \leftarrow r_{\tau,h} + \widetilde{V}_{t,h+1}(s_{\tau,h+1})$
    Fit $\hat{Q}_{t,h} \in \arg\min_{Q \in \mathcal{R}_h} \sum_{\tau=1}^{t-1} \big(Q(s_{\tau,h}, a_{\tau,h}) - y_{\tau,h}^{(t)}\big)^2$
    Compute $b_{t,h}(s,a) \leftarrow \min\big\{B_h, \beta\, U_{\mathcal{R}_h}\big(\lambda, s, a; \bar{D}_{h,t-1}^{\mathrm{RL}}\big) + \zeta_{\mathrm{RL}}\big\}$
    Set $\widetilde{Q}_{t,h}(s,a) \leftarrow \Pi_{[0,B_h]}\big(\hat{Q}_{t,h}(s,a) + b_{t,h}(s,a)\big)$
    Update policy $\pi_{t,h}(a \mid s) \propto \pi_{\mathrm{ref},h}(a \mid s) \exp\big(\eta \widetilde{Q}_{t,h}(s,a)\big)$
    Update $\widetilde{V}_{t,h}(s) \leftarrow V_h(\widetilde{Q}_{t,h}; s)$
  end for
  Observe initial state $s_{t,1}$
  for $h = 1, \dots, H$ **do**
    Choose action $a_{t,h} \sim \pi_{t,h}(\cdot \mid s_{t,h})$ and observe $(r_{t,h}, s_{t,h+1})$
  end for
end for

stagewise uncertainty bonus from $\bar{D}_{h,t-1}^{\mathrm{RL}}$ using $U_{\mathcal{R}_h}$ (Definition 4.1), forms an optimistic clipped Q-estimate, and then performs Gibbs policy improvement relative to $\pi_{\mathrm{ref},h}$. The soft value update uses the stagewise log-partition operator $V_h(\cdot\,; s)$ already defined in Section 3. All update formulas are given once in Algorithm 2, and are not repeated in the prose. For the RL guarantee in Corollary 5.4, we use $\beta = \Theta(\sqrt{\Lambda_{\mathrm{RL}}})$, where Λ_{RL} includes the finite stagewise function classes and the finite continuation-value classes used in the frozen-target concentration step. The order-level calibration is stated in Corollary 5.4, and the constant-level version is given in Appendix B.

Algorithm 2 uses labels recomputed within the current backward pass and inserts the known misspecification level ζ_{RL} directly into the stagewise bonus. Relative to classical LSVI-UCB-style methods, the key differences are the KL-regularized soft continuation value and Gibbs policy improvement.

Remark 4.2 (Relation to realizable KL-regularized RL)**.** *When $\zeta_{\mathrm{RL}} = 0$ and the stagewise approximation is exact, Algorithm 2 reduces to the standard KL-regularized optimistic soft-planning template: backward least-squares fitting, optimism at the Q-function level, and Gibbs policy improvement relative to the reference policy; see Zhao et al. (2025).*

Remark 4.3 (Unknown misspecification parameters)**.** *Algorithms 1 and 2 are written as calibrated base learners, where the bonuses use ζ_{Bandit} or ζ_{RL}. When the misspecification level is unknown, this calibration can be removed by a standard model-selection wrapper over a geometric grid of candidate radii, following the meta-algorithmic approach of Li & Yang (2024). In our setting, each base learner is simply Algorithm 1 or Algorithm 2 run with one candidate value of ζ. For any candidate radius that upper bounds the true misspecification level, the confidence and optimism arguments in Appendix B apply unchanged to that base learner. Thus unknown misspecification affects only the outer model-selection overhead, while the calibrated regret bounds below describe the base guarantees used by the wrapper.*

5 Analysis

5.1 KL-Regularized Contextual Bandits

Theorem 5.1 (KL bandit regret under misspecification)**.** *Assume the bandit model (1) with conditionally 1-sub-Gaussian noise. Assume the function class is finite with $N_{\mathcal{R}} := |\mathcal{R}| < \infty$. Let ζ_{Bandit} denote the pointwise misspecification level defined in (6). Run Algorithm 1 with the bonus calibrated to the known misspecification level, and choose $\beta = \Theta\big(\sqrt{\log(N_{\mathcal{R}} T/\delta)}\big)$. Choose λ large enough for the appendix localization step; in particular, it suffices that $\lambda \gtrsim$*

$d(\mathcal{R}, \lambda, T) + \log(N_{\mathcal{R}}T/\delta) + T\zeta^2_{\mathrm{Bandit}}$. *Then, with probability at least* $1-\delta$,

$$\mathrm{Reg}(T) = O\Big(\eta\Big[\log\Big(N_{\mathcal{R}}T/\delta\Big)\, d(\mathcal{R}, \lambda, T) + T\zeta^2_{\mathrm{Bandit}}\Big]\Big)\,. \tag{22}$$

Proof sketch. The proof is an optimistic regression analysis adapted to the KL-regularized objective, following the standard second-order-bonus route with KL geometry made explicit via the Gibbs variational identity.

The starting point is an exact Gibbs representation of the round-t KL regret gap. Under optimism, this gap is upper bounded by η times a squared score residual under the learner policy. Hence, the regret analysis reduces to controlling a second-order quantity, namely a predictable sum of squared bonuses.

To establish optimism under misspecification, we compare the ERM estimator to a comparator $R^\circ \in \mathcal{R}$ satisfying $\sup_{x\in\mathcal{X}} \sup_{a\in\mathcal{A}} |R^\circ(x,a) - R^\star(x,a)| \le \zeta_{\mathrm{Bandit}}$. The ERM localization step controls $\widehat{R}_{t-1} - R^\circ$, while (6) controls $R^\circ - R^\star$. After adding the bonus and clipping, these combine into a pointwise residual bound $0 \le \widetilde{R}_{t-1}(x,a) - R^\star(x,a) \lesssim b_{t-1}(x,a)$.

The remainder separates concentration from complexity control. A Freedman step relates the predictable squared-bonus sum to the squared-bonus sum along sampled actions, and eluder dimension summability bounds $\sum_{t=1}^{T} b_{t-1}(x_t,a_t)^2$ by $\beta^2 d(\mathcal{R}, \lambda, T)$. Finally, the additive misspecification term in the bonus contributes $T\zeta^2_{\mathrm{Bandit}}$.

Remark 5.2. *For reference, we recall the realizable high-probability KL-bandit guarantee of Zhao et al. (2025). In the realizable setting with a finite function class, their theorem implies*

$$\mathrm{Reg}(T) = O\Big(\eta\, \log\Big(N_{\mathcal{R}}T/\delta\Big)\, d(\mathcal{R}, \lambda, T)\Big)\,, \tag{23}$$

up to universal constants and lower-order logarithmic factors.

In the realizable regime $\zeta_{\mathrm{Bandit}} = 0$*, Theorem 5.1 recovers* (23)*. Relative to the realizable analysis, the misspecified extension appears only through the explicit additive term* $T\zeta^2_{\mathrm{Bandit}}$ *in* (22)*.*

5.2 KL-Regularized RL

For episodic RL, the proof follows the same high-level route as in the bandit case, but the stagewise regression problem uses recomputed targets. The guarantee below is for the known-misspecification setting used by Algorithm 2. For the frozen-target confidence step, let $\mathcal{V}_{h+1}$ denote a finite class containing the continuation values $\widetilde{V}_{t,h+1}$ generated by the algorithm. This class is used only for the proof of the frozen-target concentration event and is not an additional algorithmic input.

We first state a modular regret theorem that separates the KL-RL regret reduction from the statistical verification of confidence and uncertainty conditions. The concrete guarantee for Algorithm 2 is then obtained as a corollary by verifying these conditions in Appendix B.

Theorem 5.3 (Modular KL-RL regret under confidence and uncertainty control)**.** *Consider any optimistic KL-regularized fitted-Q procedure that produces scores* $\widetilde{Q}_{t,h}$*, policies* $\pi_{t,h}$*, and bonuses* $b_{t,h}$*. Let* $m_{t,h}(s,a) := (\mathcal{T}_{\eta,h}\widetilde{Q}_{t,h+1})(s,a)$*. Assume that with probability at least* $1-\delta$*, the following conditions hold simultaneously.*

1. ***Confidence and optimism.*** *For all* t, h, s, a*,*

$$\widetilde{Q}_{t,h}(s,a) \ge Q^\star_h(s,a), \qquad 0 \le \widetilde{Q}_{t,h}(s,a) - m_{t,h}(s,a) \le 2b_{t,h}(s,a). \tag{24}$$

2. ***Uncertainty summability.*** *There exists a universal constant* $C_b > 0$ *such that*

$$\sum_{t=1}^{T}\sum_{h=1}^{H} b_{t,h}(s_{t,h},a_{t,h})^2 \le C_b\Big(\beta^2 d_{\mathrm{RL}}(\lambda, T) + HT\zeta^2_{\mathrm{RL}}\Big). \tag{25}$$

3. ***Predictable-to-realized alignment.*** *Let* $X_{t,h} := b_{t,h}(s_{t,h}, a_{t,h})^2$ *and* $\bar{X}_{t,h} := \mathbb{E}[X_{t,h} \mid \mathcal{F}^-_{t,h}]$. *Then*

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \bar{X}_{t,h} \le 2\sum_{t=1}^{T}\sum_{h=1}^{H} X_{t,h} + 4H^2 \log\left(\frac{8}{\delta}\right). \tag{26}$$

Moreover, let $R_t := \mathrm{Reg}^{\mathrm{RL}}_{\eta}(t)$ *and* $\bar{R}_t := \mathbb{E}[R_t \mid \mathcal{F}^-_{t,1}]$. *Then*

$$\sum_{t=1}^{T} R_t \le 2\sum_{t=1}^{T} \bar{R}_t + 8H^2 \log\left(\frac{8}{\delta}\right). \tag{27}$$

Then, on the same event,

$$\mathrm{Reg}(T) = \widetilde{O}\big(\eta H^2 \beta^2 d_{\mathrm{RL}}(\lambda, T) + \eta H^3 T \zeta^2_{\mathrm{RL}}\big), \tag{28}$$

where the hidden logarithmic factors come from the confidence, union-bound, and alignment events.

Corollary 5.4 (KL-RL regret under misspecification)**.** *Assume the episodic KL-regularized RL setting in Section 3 with conditionally* 1*-sub-Gaussian reward noise, and assume the stagewise pointwise misspecification condition* (16) *with level* ζ_{RL}. *Assume each stagewise function class is finite with* $N_{\mathcal{R}_h} := |\mathcal{R}_h| < \infty$. *For the frozen-target concentration step, let* $\mathcal{V}_{h+1}$ *be the finite continuation-value class used in Appendix B, and write* $N_{\mathcal{V}_{h+1}} := |\mathcal{V}_{h+1}|$.

Run Algorithm 2 with the bonus calibrated to ζ_{RL}. *Define*

$$\Lambda_{\mathrm{RL}} := \max_{h\in[H]} \log\left(\frac{4THN_{\mathcal{R}_h}N_{\mathcal{V}_{h+1}}}{\delta}\right), \tag{29}$$

and choose

$$\beta = \Theta\Big(\sqrt{\Lambda_{\mathrm{RL}}}\Big), \qquad \lambda \gtrsim H^2 + \bar{\sigma}^2\Big(d_{\mathrm{RL}}(\lambda, T) + \Lambda_{\mathrm{RL}}\Big) + T\zeta^2_{\mathrm{RL}}, \tag{30}$$

where $\bar{\sigma}^2 = \Theta(1 + H^2)$. *Then, with probability at least* $1 - \delta$,

$$\mathrm{Reg}(T) = \widetilde{O}\big(\eta H^2 d_{\mathrm{RL}}(\lambda, T) + \eta H^3 T\, \zeta^2_{\mathrm{RL}}\big). \tag{31}$$

Remark 5.5. *We state Theorem 5.1 and Corollary 5.4 for finite function classes only to keep the high-probability confidence events simple, following Zhao et al. (2025). The uncertainty definitions in Definition 4.1 do not require finiteness. For infinite classes, the finite union bounds can be replaced by standard covering-number arguments for the corresponding localized regression classes, changing only logarithmic factors. The dependence on* $d(\mathcal{R}, \lambda, T)$, $d_{\mathrm{RL}}(\lambda, T)$, *and the misspecification levels remains the same up to these logarithmic replacements.*

Proof sketch. Theorem 5.3 gives the regret reduction once confidence, residual control, uncertainty-summability, and alignment are available. Appendix B verifies these conditions for Algorithm 2; in particular, the frozen-target confidence step is proved uniformly over $V \in \mathcal{V}_{h+1}$, which allows us to instantiate it with the data-dependent continuation value $\widetilde{V}_{t,h+1}$.

At a high level, fix an episode and condition on the start-of-episode filtration so that the planning outputs and recomputed targets are fixed. Under optimism, each stagewise Gibbs KL term is controlled by a squared Q-gap under the learner policy. A multi-step reduction propagates these stagewise terms through the episode and upper bounds the total by an H^2-weighted sum of squared Bellman residuals, which is the source of the leading H^2 factor.

Next, the Bellman residuals are controlled by the bonus. The frozen-target confidence step controls the estimation error around a stagewise comparator, while (16) controls the comparator-to-target gap. After adding the optimistic bonus and clipping, these yield a residual bound of the form $e_{t,h}(s,a)^2 \lesssim b_{t,h}(s,a)^2$, hence conditional regret reduces to a second-order bonus sum.

Finally, uncertainty-squared summability yields the leading complexity term $\eta H^2 \beta^2 d_{\mathrm{RL}}(\lambda, T)$, plus the explicit misspecification contribution $\eta H^3 T \zeta_{\mathrm{RL}}^2$. The logarithmic factors hidden in $\widetilde{O}(\cdot)$ include the finite-class union bound over $\mathcal{R}_h$, the continuation-value classes $\mathcal{V}_{h+1}$, and the Freedman alignment steps.

Remark 5.6. *For reference, we recall the realizable high-probability KL-RL guarantee of Zhao et al. (2025). In their realizable setting with a finite function class, their theorem implies*

$$\mathrm{Reg}(T) = \tilde{O}\big(\eta H^2 \, d(\mathcal{F}, \lambda, T)\big) \,, \tag{32}$$

where $d(\mathcal{F}, \lambda, T)$ denotes their eluder-dimension complexity term.

In the realizable regime $\zeta_{\mathrm{RL}} = 0$, Corollary 5.4 recovers the same leading $\eta H^2 \times$ (complexity) dependence. The misspecified extension is captured by the explicit additive term $\eta H^3 T \, \zeta_{\mathrm{RL}}^2$ in (31).

6 Conclusion

We study KL-regularized contextual bandits and episodic reinforcement learning under general function approximation with model misspecification. We formulate KL-aligned misspecification conditions and analyze optimistic regression-based algorithms that act via reference-relative Gibbs policies. For contextual bandits, we prove high-probability KL-regret bounds with explicit dependence on misspecification and localized eluder-dimension complexity. For episodic RL, we establish a high-probability regret guarantee for our KL-regularized LSVI-style algorithm, where the leading term is governed by an explicit confidence/uncertainty interface and the bound again separates statistical complexity from misspecification. Both results recover the corresponding realizable KL-regularized guarantees as special cases. Technically, the analysis combines Gibbs variational identities, optimism-based KL self-bounding, and summability of squared bonus terms, leading to logarithmic-in-T high-probability factors.

Our misspecification model is pointwise and paired with light-tailed noise assumptions. Recent RLHF studies suggest that KL regularization alone may not ensure robustness under broader forms of reward misspecification, and can fail in particular under heavy-tailed reward errors. Extending KL-regret guarantees to on-policy misspecification measures and heavy-tailed settings remains an important direction for future work. Another open question is whether the horizon dependence in the RL guarantee can be improved, ideally with matching lower bounds under KL-adapted misspecification. More broadly, a natural future direction is to extend the framework beyond reward-based learning to richer feedback models, including preference-based formulations.

Acknowledgements

We thank the anonymous reviewers and area chair for their helpful comments. QG is supported in part by the National Science Foundation DMS-2323113 and IIS-2403400. HH and HW are supported in part by National Science Foundation IIS-2403401. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert E. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, 2014.

Philip Amortila, Tongyi Cao, and Akshay Krishnamurthy. Mitigating covariate shift in misspecified regression with applications to reinforcement learning. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, 2024. URL https://proceedings.mlr.press/v247/amortila24a.html.

Shicong Cen, Jincheng Mei, Katayoon Goshvadi, Hanjun Dai, Tong Yang, Sherry Yang, Dale Schuurmans, Yuejie Chi, and Bo Dai. Value-incentivized preference optimization: A unified approach to online and offline RLHF. *arXiv preprint arXiv:2405.19320*, 2024. URL https://arxiv.org/abs/2405.19320.

Dylan Foster, Alekh Agarwal, Miroslav Dudík, Haipeng Luo, and Robert E. Schapire. Practical contextual bandits with regression oracles. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 2018. URL https://proceedings.mlr.press/v80/foster18a.html.

Dylan Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, 2021a. URL https://proceedings.mlr.press/v134/foster21a.html.

Dylan J. Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 2020. URL https://proceedings.mlr.press/v119/foster20a.html.

Dylan J. Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. *arXiv preprint arXiv:2107.05745*, 2021b. URL https://arxiv.org/abs/2107.05745.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *COLT*, 2020.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2102.00815.

Sanath Kumar Krishnamurthy, Vitor Hadad, and Susan Athey. Adapting to misspecification in contextual bandits with offline regression oracles. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 2021. URL https://proceedings.mlr.press/v139/krishnamurthy21a.html.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems*, volume 20, 2008.

Junghyun Lee, Minju Hong, Kwang-Sung Jun, Chulhee Yun, and Se-Young Yun. Regularized online RLHF with generalized bilinear preferences. *arXiv preprint arXiv:2602.23116*, 2026. URL https://arxiv.org/abs/2602.23116.

Orin Levy and Yishay Mansour. Optimal regret for policy optimization in contextual bandits. *arXiv preprint arXiv:2602.13700*, 2026. URL https://arxiv.org/abs/2602.13700.

Orin Levy, Liad Erez, Alon Peled-Cohen, and Yishay Mansour. Regret bounds for adversarial contextual bandits with general function approximation and delayed feedback. *arXiv preprint arXiv:2510.09127*, 2025. URL https://arxiv.org/abs/2510.09127.

Yunfan Li and Lin Yang. On the model-misspecification in reinforcement learning. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, 2024. URL https://proceedings.mlr.press/v238/li24m.html.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J. Mankowitz, Doina Precup, and

Bilal Piot. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023. URL https://arxiv.org/abs/2312.00886.

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. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Aldo Pacchiano, Aadirupa Saha, and Jason Lee. Dueling RL: Reinforcement learning with trajectory preferences. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, 2023. URL https://arxiv.org/abs/2111.04850.

Hao Qin and Chicheng Zhang. Taming the monster every context: Complexity measure and unified framework for offline-oracle efficient contextual bandits. *arXiv preprint arXiv:2602.09456*, 2026. URL https://arxiv.org/abs/2602.09456.

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. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.

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

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Ayano Takemura, Shinji Ito, and Junya Honda. A parameter-free algorithm for misspecified linear contextual bandits. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, 2021. URL https://proceedings.mlr.press/v130/takemura21a.html.

Daniel Vial, Advait Parulekar, Sanjay Shakkottai, and R. Srikant. Improved algorithms for misspecified linear markov decision processes. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, 2022. URL https://proceedings.mlr.press/v151/vial22a.html.

Ruosong Wang, Ruslan Salakhutdinov, and Lin F. Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2005.10804.

Zihan Wang, Boyi Liu, and Chi Jin. Is RLHF more difficult than standard RL? a theoretical analysis and an empirical study. *arXiv preprint arXiv:2306.14111*, 2023. URL https://arxiv.org/abs/2306.14111.

Di Wu, Chengshuai Shi, Jing Yang, and Cong Shen. Greedy sampling is provably efficient for RLHF. *arXiv preprint arXiv:2510.24700*, 2025. URL https://arxiv.org/abs/2510.24700.

Tianhao Wu and Wen Sun. Making reinforcement learning from human feedback efficient via randomization. *arXiv preprint arXiv:2310.14554*, 2023. URL https://arxiv.org/abs/2310.14554.

Tengyang Xie, Dylan J. Foster, Yu Bai, Nan Jiang, and Sham M. Kakade. The role of coverage in online reinforcement learning. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.04157.

Tengyang Xie, Dylan J. Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit $Q^\star$-approximation for sample-efficient RLHF. *arXiv preprint arXiv:2405.21046*, 2024. URL https://arxiv.org/abs/2405.21046.

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. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024. URL https://proceedings.mlr.press/v235/xiong24a.html.

Yaqi Xu, Zeyu Gao, Yijie Guo, Yifei Wang, and Zhuoran Yang. Preference-based reinforcement learning with finite-time guarantees. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2006.08910.

Chenlu Ye, Wei Xiong, Yuheng Zhang, Hanze Dong, Nan Jiang, and Tong Zhang. Online iterative reinforcement learning from human feedback with general preference model. *arXiv preprint arXiv:2402.07314*, 2024. URL https://arxiv.org/abs/2402.07314.

Tong Zhang. *Mathematical Analysis of Machine Learning Algorithms*. Cambridge University Press, Cambridge, 2023. DOI: 10.1017/9781009093057.

Heyang Zhao, Chenlu Ye, Quanquan Gu, and Tong Zhang. Sharp analysis for KL-regularized contextual bandits and RLHF. *arXiv preprint arXiv:2411.04625*, 2024. URL https://arxiv.org/abs/2411.04625.

Heyang Zhao, Chenlu Ye, Wei Xiong, Quanquan Gu, and Tong Zhang. Logarithmic regret for online KL-regularized reinforcement learning. *arXiv preprint arXiv:2502.07460*, 2025. URL https://arxiv.org/abs/2502.07460.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Notation

Symbol	Meaning
T, H	Number of rounds/episodes and episodic horizon length.
δ	Target failure probability in high-probability guarantees.
η	KL-regularization parameter.
λ, β	Localization parameter and optimism-bonus scale.
π_{ref}, $\pi_{\mathrm{ref},h}$	Reference policy in contextual bandits and stage-h reference policy in RL.
$J(\pi)$, $\mathrm{Reg}(T)$	KL-regularized objective and cumulative KL regret.
$R^\star$, $r_h^\star$	Ground-truth bandit reward and stage-h RL reward.
$\mathcal{R}$, $\mathcal{R}_h$	Bandit reward-function class and stage-h RL function class.
$N_{\mathcal{R}}$, $N_{\mathcal{R}_h}$	Cardinalities of $\mathcal{R}$ and $\mathcal{R}_h$.
ζ_{Bandit}, ζ_{RL}	Bandit and RL misspecification levels.
B_h	Stagewise value range upper bound, $B_h := H - h + 1$.
$Q_h^\star, V_h^\star$	Optimal KL-regularized Q-function and value function at stage h.
$\mathcal{T}_{\eta,h}$	Stage-h KL-regularized Bellman operator.
$V_h(f; s)$	Stage-h log-partition, or soft-value, operator induced by score f.
$\widehat{Q}_{t,h}$, $\widetilde{Q}_{t,h}$	Stage-h fitted Q-estimate and optimistic clipped score in episode t.
$\widetilde{V}_{t,h}$	Soft value induced by $\widetilde{Q}_{t,h}$.
$m_{t,h}$	Bellman target mean $m_{t,h} := \mathcal{T}_{\eta,h}\widetilde{Q}_{t,h+1}$.
$b_{t,h}(s,a)$	Stage-h RL optimism bonus for (s,a) in episode t.
$\bar{D}_{h,t-1}^{\mathrm{RL}}$	Stage-h state-action history before episode t.
$\mathcal{V}_{h+1}$, $N_{\mathcal{V}_{h+1}}$	Finite continuation-value class used in the frozen-target concentration step and its cardinality.
$d(\mathcal{R}, \lambda, T)$	Eluder-style complexity of the bandit function class.
$d(\mathcal{R}_h, \lambda, T)$	Stagewise RL eluder-style complexity.
$d_{\mathrm{RL}}(\lambda, T)$	Aggregate RL complexity, $d_{\mathrm{RL}}(\lambda, T) := \sum_{h=1}^{H} d(\mathcal{R}_h, \lambda, T)$.
Λ_{RL}	Logarithmic union-bound factor for the RL finite-class concentration argument.
$\mathrm{unc}_{t,h}(s,a)$	Stagewise localized uncertainty width used to define $b_{t,h}$.
$X_{t,h}$, $\bar{X}_{t,h}$	Realized bonus square and its predictable counterpart in the RL alignment step.
R_t, $\bar{R}_t$	Episode-t realized KL-regret contribution and its conditional expectation.
$\mathcal{F}_t^-$, $\mathcal{F}_{t,h}^-$	Pre-action filtrations in bandits and episodic RL.
$\bar{\sigma}^2$	Uniform sub-Gaussian proxy for recomputed RL regression labels.

B Proofs

B.1 KL-Regularized Contextual Bandits

Define the bandit filtrations

$$\mathcal{F}_t^- := \sigma\big(\{(x_i, a_i, r_i)\}_{i=1}^{t-1}, x_t\big), \qquad \mathcal{F}_t^a := \sigma(\mathcal{F}_t^-, a_t), \qquad \mathcal{F}_t := \sigma(\mathcal{F}_t^a, r_t).$$

For alignment arguments, we view $X_t - \mathbb{E}[X_t \mid \mathcal{F}_t^-]$ as a martingale difference with respect to the shifted filtration $\{\mathcal{F}_t^-\}_{t\geq 1}$, since X_t is $\mathcal{F}_{t+1}^-$-measurable and $\mathbb{E}[X_t \mid \mathcal{F}_t^-]$ is predictable.

Lemma B.1 (Variational form and Gibbs optimizer)**.** *Fix a context x and any measurable score $R(x,\cdot)$ such that $\exp(\eta R(x,\cdot))$ is integrable under $\pi_{\mathrm{ref}}(\cdot \mid x)$. Define*

$$\mathcal{L}_\eta(x; R) = \frac{1}{\eta} \log \mathbb{E}_{a\sim\pi_{\mathrm{ref}}(\cdot|x)}\Big[\exp\big(\eta R(x,a)\big)\Big]. \tag{33}$$

For any $\pi(\cdot \mid x) \ll \pi_{\mathrm{ref}}(\cdot \mid x)$ *define*

$$\mathcal{J}_\eta(\pi; x, R) := \mathbb{E}_{a\sim\pi(\cdot|x)}[R(x,a)] - \frac{1}{\eta}\mathrm{KL}(\pi(\cdot \mid x) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid x)). \tag{34}$$

Then

$$\mathcal{L}_\eta(x; R) = \max_{\pi(\cdot|x)\ll\pi_{\mathrm{ref}}(\cdot|x)} \mathcal{J}_\eta(\pi; x, R), \tag{35}$$

and the unique maximizer is the Gibbs policy

$$\pi_R(a \mid x) = \frac{\pi_{\mathrm{ref}}(a \mid x)\exp(\eta R(x,a))}{\mathbb{E}_{a'\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\eta R(x,a'))]}. \tag{36}$$

Moreover, for any $\pi(\cdot \mid x) \ll \pi_{\mathrm{ref}}(\cdot \mid x)$*,*

$$\mathcal{L}_\eta(x; R) - \mathcal{J}_\eta(\pi; x, R) = \frac{1}{\eta}\mathrm{KL}(\pi(\cdot \mid x) \,\|\, \pi_R(\cdot \mid x)). \tag{37}$$

Proof. Fix x and abbreviate $\pi_{\mathrm{ref}}(\cdot \mid x)$ by π_{ref}. For any $\pi \ll \pi_{\mathrm{ref}}$, let $w(a) := \frac{d\pi}{d\pi_{\mathrm{ref}}}(a)$ so that $\mathbb{E}_{\pi_{\mathrm{ref}}}[w] = 1$. Then $\mathrm{KL}(\pi\|\pi_{\mathrm{ref}}) = \mathbb{E}_{\pi_{\mathrm{ref}}}[w \log w]$, and

$$\mathcal{J}_\eta(\pi; x, R) = \mathbb{E}_{\pi_{\mathrm{ref}}}\left[w(a)R(x,a) - \frac{1}{\eta}w(a)\log w(a)\right].$$

Maximizing over $w \geq 0$ subject to $\mathbb{E}_{\pi_{\mathrm{ref}}}[w] = 1$ yields $w(a) \propto e^{\eta R(x,a)}$, which gives (36). Plugging this optimizer into $\mathcal{J}_\eta$ yields (33)–(35).

Finally, since $\pi_R(a) \propto \pi_{\mathrm{ref}}(a)e^{\eta R(x,a)}$,

$$\begin{aligned}\mathrm{KL}(\pi\|\pi_R) &= \mathbb{E}_\pi\left[\log\frac{\pi}{\pi_{\mathrm{ref}}} - \eta R(x,a) + \log\mathbb{E}_{\pi_{\mathrm{ref}}}[e^{\eta R(x,\cdot)}]\right] \\ &= \eta\Big(\mathcal{L}_\eta(x; R) - \mathcal{J}_\eta(\pi; x, R)\Big),\end{aligned}$$

which is (37). □

Lemma B.2 (Bandit optimism implies quadratic self-bounding of Gibbs KL)**.** *Fix a context* x*. Let* $u(\cdot), v(\cdot)$ *be scores on* $\mathcal{A}(x)$ *such that* $u(a) \geq v(a)$ *for all* $a \in \mathcal{A}(x)$*. Let* π_u, π_v *be the induced Gibbs policies w.r.t.* $\pi_{\mathrm{ref}}(\cdot \mid x)$*. Then for any* $\eta > 0$*,*

$$\frac{1}{\eta}\mathrm{KL}\big(\pi_u(\cdot \mid x) \,\|\, \pi_v(\cdot \mid x)\big) \leq \eta\,\mathbb{E}_{a\sim\pi_u(\cdot|x)}\big[(u(a) - v(a))^2\big]. \tag{38}$$

Proof. Let $F(w) := \log\mathbb{E}_{a\sim\pi_{\mathrm{ref}}(\cdot|x)}[e^{\eta w(a)}]$ and $\Delta := u - v \geq 0$. As before, $\mathrm{KL}(\pi_u\|\pi_v)$ is the Bregman divergence of F, and with $w_\lambda := v + \lambda\Delta$,

$$\mathrm{KL}(\pi_u\|\pi_v) \leq \eta^2\int_0^1 (1-\lambda)\mathbb{E}_{\pi_{w_\lambda}}[\Delta^2]\,d\lambda.$$

Define $\phi(\lambda) := \mathbb{E}_{\pi_{w_\lambda}}[\Delta^2]$. Differentiating Gibbs expectations yields

$$\phi'(\lambda) = \eta\ \mathrm{Cov}_{a\sim\pi_{w_\lambda}}(\Delta(a)^2, \Delta(a)) \geq 0,$$

since Δ^2 is a nondecreasing function of Δ on $[0,\infty)$. Hence ϕ is nondecreasing, so $\mathbb{E}_{\pi_{w_\lambda}}[\Delta^2] \leq \mathbb{E}_{\pi_u}[\Delta^2]$ for all $\lambda \in [0,1]$. Therefore,

$$\mathrm{KL}(\pi_u\|\pi_v) \leq \eta^2\Big(\int_0^1 (1-\lambda)\,d\lambda\Big)\mathbb{E}_{\pi_u}[\Delta^2] = \frac{\eta^2}{2}\mathbb{E}_{\pi_u}[\Delta^2] \leq \eta^2\mathbb{E}_{\pi_u}[\Delta^2].$$

Dividing by η gives (38). □

Using the main-text bandit history notation, let $z_i := (x_i, a_i)$. For any $\bar{D}_{t-1}^{\mathrm{Bandit}}$, define the localized uncertainty width

$$\mathcal{U}_{\mathcal{R}}\big(\lambda; (x,a) \mid \bar{D}_{t-1}^{\mathrm{Bandit}}\big) := \min\left\{1, \sup_{\substack{R,R' \in \mathcal{R}: \\ \sum_{i=1}^{t-1}(R(z_i)-R'(z_i))^2 \le \lambda}} \big|R(x,a) - R'(x,a)\big|\right\}. \tag{39}$$

Fix $\delta \in (0,1)$ and set

$$\beta := \max\Big\{1,\ c_0\sqrt{\log\Big(\frac{2TN_{\mathcal{R}}}{\delta}\Big)}\Big\}, \qquad \lambda := c_\lambda\Big(d(\mathcal{R},\lambda,T) + \log\tfrac{2TN_{\mathcal{R}}}{\delta} + T\,\zeta_{\mathrm{Bandit}}^2\Big), \tag{40}$$

for sufficiently large universal constants. The corresponding bonus is

$$b_{t-1}(x,a) := \min\left\{1,\ \beta\,\mathcal{U}_{\mathcal{R}}\big(\lambda; (x,a) \mid \bar{D}_{t-1}^{\mathrm{Bandit}}\big) + \zeta_{\mathrm{Bandit}}\right\}. \tag{41}$$

Lemma B.3 (Finite-class offset inequality for bandit regression)**.** *Assume the noise ϵ_t in (1) is conditionally zero-mean and 1-sub-Gaussian w.r.t. $\mathcal{F}_t^a$. Fix any deterministic comparator $R^\dagger \in \mathcal{R}$. Then there exists an event $\mathcal{E}_1$ with $\mathbb{P}(\mathcal{E}_1) \ge 1 - \delta/2$ such that, on $\mathcal{E}_1$, simultaneously for all $t \in [T]$ and all $R \in \mathcal{R}$,*

$$\sum_{i=1}^{t-1} \epsilon_i\big(R(z_i) - R^\dagger(z_i)\big) \le \frac{1}{8}\sum_{i=1}^{t-1}\big(R(z_i) - R^\dagger(z_i)\big)^2 + 4\log\left(\frac{2TN_{\mathcal{R}}}{\delta}\right). \tag{42}$$

Proof. Fix $R \in \mathcal{R}$ and define

$$g_i^R := R(z_i) - R^\dagger(z_i).$$

Since $z_i = (x_i, a_i)$ is observed before the reward noise ϵ_i, g_i^R is $\mathcal{F}_i^a$-measurable. For any $\alpha > 0$, conditional sub-Gaussianity gives

$$\mathbb{E}\left[\exp\left(\alpha\epsilon_i g_i^R - \frac{\alpha^2}{2}(g_i^R)^2\right) \,\middle|\, \mathcal{F}_i^a\right] \le 1. \tag{43}$$

Iterating (43) over $i = 1, \dots, t-1$, applying Markov's inequality, and taking a union bound over $t \in [T]$ and $R \in \mathcal{R}$, we obtain that with probability at least $1 - \delta/2$,

$$\alpha\sum_{i=1}^{t-1}\epsilon_i g_i^R - \frac{\alpha^2}{2}\sum_{i=1}^{t-1}(g_i^R)^2 \le \log\left(\frac{2TN_{\mathcal{R}}}{\delta}\right) \tag{44}$$

holds simultaneously for all t and R. Choosing $\alpha = 1/4$ yields (42). □

Lemma B.4 (Misspecified ERM localization for bandits)**.** *Let $\hat{R}_{t-1} \in \arg\min_{R\in\mathcal{R}} \sum_{i=1}^{t-1}(R(z_i) - r_i)^2$ be the ERM. Let $R^\dagger \in \mathcal{R}$ be a fixed comparator satisfying*

$$\sup_{x\in\mathcal{X}}\sup_{a\in\mathcal{A}} |R^\dagger(x,a) - R^\star(x,a)| \le \zeta_{\mathrm{Bandit}}. \tag{45}$$

Then, on $\mathcal{E}_1$, simultaneously for all $t \in [T]$,

$$\sum_{i=1}^{t-1}\big(\hat{R}_{t-1}(z_i) - R^\dagger(z_i)\big)^2 \le C_{\mathrm{loc}}\left[\log\left(\frac{2TN_{\mathcal{R}}}{\delta}\right) + T\zeta_{\mathrm{Bandit}}^2\right], \tag{46}$$

for a universal constant $C_{\mathrm{loc}} > 0$. In particular, the choice (40) with c_λ large enough implies

$$\sum_{i=1}^{t-1}\big(\hat{R}_{t-1}(z_i) - R^\dagger(z_i)\big)^2 \le \lambda \tag{47}$$

simultaneously for all $t \in [T]$.

Proof. Fix $t \in [T]$ and abbreviate

$$\Delta_i := \hat{R}_{t-1}(z_i) - R^\dagger(z_i), \qquad g_i := R^\dagger(z_i) - R^\star(z_i).$$

By ERM optimality,

$$\sum_{i=1}^{t-1} \big(\hat{R}_{t-1}(z_i) - r_i\big)^2 \le \sum_{i=1}^{t-1} \big(R^\dagger(z_i) - r_i\big)^2. \tag{48}$$

Using $r_i = R^\star(z_i) + \epsilon_i$, we have

$$R^\dagger(z_i) - r_i = g_i - \epsilon_i, \qquad \hat{R}_{t-1}(z_i) - r_i = \Delta_i + g_i - \epsilon_i.$$

Expanding (48) and cancelling common terms gives

$$\sum_{i=1}^{t-1} \Delta_i^2 \le 2\sum_{i=1}^{t-1} \epsilon_i \Delta_i - 2\sum_{i=1}^{t-1} g_i \Delta_i. \tag{49}$$

On $\mathcal{E}_1$, Lemma B.3 applied with $R = \hat{R}_{t-1}$ gives

$$\sum_{i=1}^{t-1} \epsilon_i \Delta_i \le \frac{1}{8}\sum_{i=1}^{t-1} \Delta_i^2 + 4\log\left(\frac{2TN_{\mathcal{R}}}{\delta}\right). \tag{50}$$

For the misspecification term, Young's inequality yields

$$2\left|\sum_{i=1}^{t-1} g_i \Delta_i\right| \le \frac{1}{4}\sum_{i=1}^{t-1} \Delta_i^2 + 4\sum_{i=1}^{t-1} g_i^2. \tag{51}$$

Combining (49), (50), and (51), we obtain

$$\sum_{i=1}^{t-1} \Delta_i^2 \le \frac{1}{2}\sum_{i=1}^{t-1} \Delta_i^2 + 8\log\left(\frac{2TN_{\mathcal{R}}}{\delta}\right) + 4\sum_{i=1}^{t-1} g_i^2. \tag{52}$$

By (45), $\sum_{i=1}^{t-1} g_i^2 \le T\zeta_{\mathrm{Bandit}}^2$. Rearranging proves (46). The final statement follows because (40) contains the right-hand side, up to a sufficiently large universal constant. □

Lemma B.5 (Uniform confidence around the best comparator (bandit))**.** *Let $R^\dagger$ be a minimizer in (6). Then on $\mathcal{E}_1$, simultaneously for all $t \in [T]$ and all (x, a),*

$$\big|\hat{R}_{t-1}(x,a) - R^\dagger(x,a)\big| \le \mathcal{U}_{\mathcal{R}}\big(\lambda; (x,a) \mid \bar{D}_{t-1}^{\mathrm{Bandit}}\big). \tag{53}$$

Proof. On $\mathcal{E}_1$, Lemma B.4 yields

$$\sum_{i=1}^{t-1} (\hat{R}_{t-1}(x_i,a_i) - R^\dagger(x_i,a_i))^2 \le \lambda$$

for all t. By the definition (39), the pair $(\hat{R}_{t-1}, R^\dagger)$ is admissible in the supremum, which gives (53). □

Lemma B.6 (Width-sum bound via eluder dimension)**.** *Let $z_t = (x_t, a_t)$ and $b_{t-1}(z_t)$ be defined in (41). Then deterministically,*

$$\sum_{t=1}^{T} b_{t-1}(z_t)^2 \le c_w\Big(\beta^2\, d(\mathcal{R}, \lambda, T) + T\zeta_{\mathrm{Bandit}}^2\Big), \tag{54}$$

for a universal constant $c_w > 0$.

Proof. Let $u_t := \mathcal{U}_{\mathcal{R}}(\lambda; z_t \mid \bar{D}_{t-1}^{\mathrm{Bandit}}) \in [0,1]$. By (41),

$$b_{t-1}(z_t) = \min\{1, \beta u_t + \zeta_{\mathrm{Bandit}}\}.$$

Using $\min\{1,a\}^2 \le a^2$ and $(u+v)^2 \le 2u^2 + 2v^2$,

$$b_{t-1}(z_t)^2 \le (\beta u_t + \zeta_{\mathrm{Bandit}})^2 \le 2\beta^2 u_t^2 + 2\zeta_{\mathrm{Bandit}}^2.$$

Summing over t gives

$$\sum_{t=1}^{T} b_{t-1}(z_t)^2 \le 2\beta^2 \sum_{t=1}^{T} u_t^2 + 2T\zeta_{\mathrm{Bandit}}^2.$$

A standard dyadic peeling + eluder counting argument yields $\sum_{t=1}^{T} u_t^2 \lesssim d(\mathcal{R}, \lambda, T)$. Absorb universal constants into c_w. □

Lemma B.7 (Freedman alignment for bandit bonus squares)**.** *Let* $X_t := b_{t-1}(x_t, a_t)^2 \in [0,1]$*, and define*

$$\bar{X}_t := \mathbb{E}[X_t \mid \mathcal{F}_t^-] = \mathbb{E}_{a\sim\pi_t(\cdot|x_t)}[b_{t-1}(x_t,a)^2]. \tag{55}$$

Then there exists an event $\mathcal{E}_3$ *with* $\mathbb{P}(\mathcal{E}_3) \ge 1-\delta/2$ *such that on* $\mathcal{E}_3$*,*

$$\sum_{t=1}^{T} \bar{X}_t \le 2\sum_{t=1}^{T} X_t + 4\log\Big(\frac{2}{\delta}\Big). \tag{56}$$

Proof. Define the shifted martingale differences $Y_{t+1} := X_t - \bar{X}_t$ for $t \in [T]$ with respect to $\{\mathcal{F}_t^-\}_{t\ge1}$. Then Y_{t+1} is $\mathcal{F}_{t+1}^-$-measurable and $\mathbb{E}[Y_{t+1} \mid \mathcal{F}_t^-] = 0$. Also $|Y_{t+1}| \le 1$, and

$$\mathbb{E}[Y_{t+1}^2 \mid \mathcal{F}_t^-] \le \mathbb{E}[X_t \mid \mathcal{F}_t^-] = \bar{X}_t.$$

Let $M_{T+1} := \sum_{t=1}^{T} Y_{t+1}$ and $V := \sum_{t=1}^{T} \mathbb{E}[Y_{t+1}^2 \mid \mathcal{F}_t^-] \le \sum_{t=1}^{T} \bar{X}_t$. Freedman's inequality applied to $-M_{T+1}$ yields, with probability at least $1-\delta/2$,

$$\sum_{t=1}^{T} (\bar{X}_t - X_t) \le \sqrt{2V\log(2/\delta)} + \tfrac{1}{3}\log(2/\delta).$$

Using $\sqrt{2V\log(2/\delta)} \le \frac{1}{2}V + \log(2/\delta)$ and rearranging gives (56) (up to slightly looser constants). □

Proof of Theorem 5.1. Work on $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_3$, which has probability at least $1-\delta$. Lemma B.6 is deterministic.

Fix t, and abbreviate $\pi_t(\cdot \mid x_t)$ by π_t. Let

$$u(a) := \widetilde{R}_{t-1}(x_t, a), \qquad v(a) := R^\star(x_t, a),$$

so that π_t is the Gibbs policy induced by u, and define

$$\pi^\star(\cdot \mid x_t) := \pi_{R^\star}(\cdot \mid x_t)$$

as the Gibbs policy induced by v.

By Lemma B.1 with $R = R^\star$,

$$\frac{1}{\eta}\mathrm{KL}(\pi_t \,\|\, \pi^\star(\cdot \mid x_t)) = \mathcal{L}_\eta(x_t; R^\star) - \mathcal{J}_\eta(\pi_t; x_t, R^\star).$$

On $\mathcal{E}_1$, Lemma B.5 gives, for all a,

$$|\hat{R}_{t-1}(x_t,a) - R^\dagger(x_t,a)| \le \mathcal{U}_{\mathcal{R}}\big(\lambda; (x_t,a) \mid \bar{D}_{t-1}^{\mathrm{Bandit}}\big).$$

By the definition of ζ_{Bandit} in (6), for all a,

$$|R^\dagger(x_t,a) - R^\star(x_t,a)| \le \zeta_{\mathrm{Bandit}}.$$

Hence, for all a,

$$|\hat{R}_{t-1}(x_t,a) - R^\star(x_t,a)| \le \mathcal{U}_{\mathcal{R}}\big(\lambda; (x_t,a) \mid \bar{D}_{t-1}^{\mathrm{Bandit}}\big) + \zeta_{\mathrm{Bandit}} \le b_{t-1}(x_t,a),$$

where the last step uses $\beta \ge 1$ and (41).

Since $\hat{R}_{t-1}(x_t,a) + b_{t-1}(x_t,a) \ge R^\star(x_t,a)$, $R^\star \in [0,1]$, and clipping onto $[0,1]$ is monotone and non-expansive, for all a,

$$\widetilde{R}_{t-1}(x_t,a) \ge R^\star(x_t,a) \qquad \text{and} \qquad 0 \le \widetilde{R}_{t-1}(x_t,a) - R^\star(x_t,a) \le 2b_{t-1}(x_t,a). \tag{57}$$

Applying Lemma B.2 with $u = \widetilde{R}_{t-1}(x_t,\cdot)$ and $v = R^\star(x_t,\cdot)$, we get

$$\frac{1}{\eta}\mathrm{KL}(\pi_t \,\|\, \pi^\star(\cdot \mid x_t)) \le \eta\, \mathbb{E}_{a\sim\pi_t}\Big[(\widetilde{R}_{t-1}(x_t,a) - R^\star(x_t,a))^2\Big].$$

Since $\pi_t \ll \pi_{\mathrm{ref}}(\cdot \mid x_t)$, the bound (57) applies π_t-a.s., hence

$$\frac{1}{\eta}\mathrm{KL}(\pi_t \,\|\, \pi^\star(\cdot \mid x_t)) \le 4\eta\, \mathbb{E}_{a\sim\pi_t}\big[b_{t-1}(x_t,a)^2\big].$$

Summing over $t \in [T]$ yields

$$\mathrm{Reg}(T) \le 4\eta \sum_{t=1}^{T} \mathbb{E}_{a\sim\pi_t}\big[b_{t-1}(x_t,a)^2\big]. \tag{58}$$

On $\mathcal{E}_3$, Lemma B.7 implies

$$\sum_{t=1}^{T} \mathbb{E}_{a\sim\pi_t}[b_{t-1}(x_t,a)^2] \le 2\sum_{t=1}^{T} b_{t-1}(x_t,a_t)^2 + 4\log\Big(\frac{2}{\delta}\Big).$$

By Lemma B.6,

$$\sum_{t=1}^{T} b_{t-1}(x_t,a_t)^2 \le c_w\Big(\beta^2 d(\mathcal{R},\lambda,T) + T\zeta_{\mathrm{Bandit}}^2\Big).$$

Substituting into (58) yields (22) up to universal constants. □

B.2 KL-Regularized RL

For episode t and stage h, let $\mathcal{F}_{t,h}^-$ be the σ-field generated by the trajectory history up to stage h (including $s_{t,h}$), and let $\mathcal{F}_{t,h}^a := \sigma(\mathcal{F}_{t,h}^-, a_{t,h})$. We use $\bar{D}_{h,t-1}^{\mathrm{RL}}$ for the stage-h state–action history. For each stage h, define the KL log-partition operator

$$V_h(u;s) := \frac{1}{\eta}\log \mathbb{E}_{a\sim\pi_{\mathrm{ref},h}(\cdot|s)}\Big[e^{\eta u(s,a)}\Big]. \tag{59}$$

We assume $\mathbb{E}[\epsilon_{t,h} \mid \mathcal{F}_{t,h}^a] = 0$ and $\epsilon_{t,h}$ is conditionally 1-sub-Gaussian given $\mathcal{F}_{t,h}^a$.

In the RL part below, we will use the conditional quantity $\bar{R}_t = \mathbb{E}[R_t \mid \mathcal{F}_{t,1}^-]$ (defined later) as a convenient intermediate. No additional concept beyond this conditionalization is intended.

Lemma B.8 (Range of the log-partition operator in RL)**.** *Fix stage h and state s. If $m \le u(s,a) \le M$ for all a in the reference-policy support, then $V_h(u;s) \in [m, M]$. In particular, if $u(s,a) \in [0, B]$ on the reference support, then $V_h(u;s) \in [0, B]$.*

Proof. If $m \le u \le M$ pointwise on the support, then $e^{\eta m} \le \mathbb{E}[e^{\eta u}] \le e^{\eta M}$, hence $m \le \frac{1}{\eta} \log \mathbb{E}[e^{\eta u}] \le M$. □

Assume the pathwise stagewise misspecification condition (16). In particular, for each algorithm-generated continuation $\widetilde{Q}_{t,h+1}$, there exists $f^\dagger_{t,h} \in \mathcal{R}_h$ such that

$$\sup_{s\in\mathcal{S}} \sup_{a\in\mathcal{A}} \left| f^\dagger_{t,h}(s,a) - (\mathcal{T}_{\eta,h}\widetilde{Q}_{t,h+1})(s,a) \right| \le \zeta_{\mathrm{RL}}. \tag{60}$$

We now define the stagewise uncertainty score directly from localized widths, so that the uncertainty-squared summability statement becomes a deterministic consequence of the width definition.

For each stage $h \in [H]$, let $z_{i,h} := (s_{i,h}, a_{i,h})$. For any $\bar{D}^{\mathrm{RL}}_{h,t-1}$, define the stagewise localized uncertainty width

$$\mathcal{U}_{\mathcal{R}_h}\big(\lambda; (s,a) \mid \bar{D}^{\mathrm{RL}}_{h,t-1}\big) := \min\left\{ B_h, \sup_{\substack{f,f'\in\mathcal{R}_h: \\ \sum_{i=1}^{t-1}(f(z_{i,h})-f'(z_{i,h}))^2 \le \lambda}} \left| f(s,a) - f'(s,a) \right| \right\}, \tag{61}$$

where $B_h := H - h + 1$.

Define the stagewise uncertainty score by

$$\mathrm{unc}_{t,h}(s,a) := \mathcal{U}_{\mathcal{R}_h}\big(\lambda; (s,a) \mid \bar{D}^{\mathrm{RL}}_{h,t-1}\big). \tag{62}$$

For the frozen-target concentration argument, take $\mathcal{V}_{h+1} \subseteq [0, B_{h+1}]^{\mathcal{S}}$ to be a finite deterministic class containing every continuation value $\widetilde{V}_{t,h+1}$ that can be generated by the algorithm at stage $h+1$, and set $N_{\mathcal{V}_{h+1}} := |\mathcal{V}_{h+1}|$. We use the convention $B_{H+1} = 0$ and $\mathcal{V}_{H+1} = \{0\}$. For each $V \in \mathcal{V}_{h+1}$, the stagewise misspecification condition gives a fixed comparator $f^{\dagger,V}_h \in \mathcal{R}_h$ satisfying

$$\sup_{s\in\mathcal{S}} \sup_{a\in\mathcal{A}} \left| f^{\dagger,V}_h(s,a) - \big(r^\star_h(s,a) + \mathbb{E}_{s'\sim P_h(\cdot|s,a)}[V(s')]\big) \right| \le \zeta_{\mathrm{RL}}. \tag{63}$$

For the realized continuation $V = \widetilde{V}_{t,h+1}$, we write $f^\dagger_{t,h} := f^{\dagger,\widetilde{V}_{t,h+1}}_h$, which is consistent with (60).

Fix $\delta \in (0,1)$ and define

$$\Lambda_{\mathrm{RL}} := \max_{h\in[H]} \log\left(\frac{4THN_{\mathcal{R}_h} N_{\mathcal{V}_{h+1}}}{\delta} \right). \tag{64}$$

Set

$$\beta := \max\Big\{1,\ c_0\sqrt{\Lambda_{\mathrm{RL}}}\Big\}, \qquad \lambda := \max\Big\{2H^2,\ c_\lambda\Big(\bar{\sigma}^2\big(d_{\mathrm{RL}}(\lambda,T) + \Lambda_{\mathrm{RL}}\big) + T\,\zeta^2_{\mathrm{RL}}\Big)\Big\}, \tag{65}$$

for sufficiently large universal constants, where $\bar{\sigma}^2 := c_\sigma(1+H^2)$. The stagewise bonus is

$$b_{t,h}(s,a) := \min\{B_h,\ \beta\,\mathrm{unc}_{t,h}(s,a) + \zeta_{\mathrm{RL}}\}. \tag{66}$$

Define the optimistic scores, soft values, and Bellman target means by

$$\widetilde{Q}_{t,h}(s,a) := \Pi[0,B_h]\big(\hat{Q}_{t,h}(s,a) + b_{t,h}(s,a)\big), \tag{67}$$

$$\widetilde{V}_{t,h}(s) := V_h(\widetilde{Q}_{t,h};s), \tag{68}$$

$$m_{t,h}(s,a) := (\mathcal{T}_{\eta,h}\widetilde{Q}_{t,h+1})(s,a) = r^\star_h(s,a) + \mathbb{E}_{s'\sim P_h(\cdot|s,a)}\big[\widetilde{V}_{t,h+1}(s')\big], \tag{69}$$

with $\widetilde{Q}_{t,H+1} \equiv 0$.

For each $t \in [T]$, stage $h \in [H]$, and past episode $i < t$, define the recomputed label

$$y_{i,h}^{(t)} := r_{i,h} + \widetilde{V}_{t,h+1}(s_{i,h+1}), \tag{70}$$

and the stagewise ERM

$$\hat{Q}_{t,h} \in \arg\min_{f \in \mathcal{R}_h} \sum_{i=1}^{t-1} \big(f(z_{i,h}) - y_{i,h}^{(t)}\big)^2. \tag{71}$$

The key technical point is that $\widetilde{V}_{t,h+1}$ is data-dependent. We therefore prove the offset inequality uniformly over $\mathcal{V}_{h+1}$ and only then instantiate it with $V = \widetilde{V}_{t,h+1}$.

Lemma B.9 (Uniform frozen-target offset inequality)**.** *There exists an event $\mathcal{E}_{\mathrm{RL}}^{\mathrm{off}}$ with $\mathbb{P}(\mathcal{E}_{\mathrm{RL}}^{\mathrm{off}}) \geq 1 - \delta/2$ such that, on this event, simultaneously for all $h \in [H]$, $t \in [T]$, $V \in \mathcal{V}_{h+1}$, and $f \in \mathcal{R}_h$, the following holds. Define*

$$m_h^V(s,a) := r_h^\star(s,a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[V(s')], \tag{72}$$

$$y_{i,h}^V := r_{i,h} + V(s_{i,h+1}), \qquad \xi_{i,h}^V := y_{i,h}^V - m_h^V(z_{i,h}). \tag{73}$$

Then

$$\sum_{i=1}^{t-1} \xi_{i,h}^V \big(f(z_{i,h}) - f_h^{\dagger,V}(z_{i,h})\big) \leq \frac{1}{8} \sum_{i=1}^{t-1} \big(f(z_{i,h}) - f_h^{\dagger,V}(z_{i,h})\big)^2 + 4\bar{\sigma}^2 \log\left(\frac{4THN_{\mathcal{R}_h} N_{\mathcal{V}_{h+1}}}{\delta}\right). \tag{74}$$

Proof. Fix h, $V \in \mathcal{V}_{h+1}$, and $f \in \mathcal{R}_h$, and define

$$g_{i,h}^{f,V} := f(z_{i,h}) - f_h^{\dagger,V}(z_{i,h}).$$

Since f, V, and $f_h^{\dagger,V}$ are fixed for this argument and $z_{i,h}$ is observed before $r_{i,h}$ and $s_{i,h+1}$, the quantity $g_{i,h}^{f,V}$ is $\mathcal{F}_{i,h}^a$-measurable.

By the definitions of $y_{i,h}^V$ and m_h^V,

$$\xi_{i,h}^V = \epsilon_{i,h} + V(s_{i,h+1}) - \mathbb{E}[V(s') \mid z_{i,h}]. \tag{75}$$

Hence $\mathbb{E}[\xi_{i,h}^V \mid \mathcal{F}_{i,h}^a] = 0$. The reward noise $\epsilon_{i,h}$ is conditionally 1-sub-Gaussian. Since $V \in [0, B_{h+1}]^{\mathcal{S}}$, Hoeffding's lemma implies that $V(s_{i,h+1}) - \mathbb{E}[V(s') \mid z_{i,h}]$ is conditionally $B_{h+1}^2/4$-sub-Gaussian. Even without conditional independence between the two noise terms, Cauchy–Schwarz gives, for every $\alpha \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}\big[\exp\big(\alpha \xi_{i,h}^V\big) \,\big|\, \mathcal{F}_{i,h}^a\big] &\leq \big(\mathbb{E}\big[\exp(2\alpha\epsilon_{i,h}) \,\big|\, \mathcal{F}_{i,h}^a\big]\big)^{1/2} \Big(\mathbb{E}\Big[\exp\big(2\alpha\big(V(s_{i,h+1}) - \mathbb{E}[V(s') \mid z_{i,h}]\big)\big) \,\Big|\, \mathcal{F}_{i,h}^a\Big]\Big)^{1/2} \\ &\leq \exp\big(c\,\alpha^2(1+H^2)\big)\end{aligned} \tag{76}$$

for a universal constant $c > 0$. Thus, after choosing $\bar{\sigma}^2 = c_\sigma(1+H^2)$ with c_σ sufficiently large, $\xi_{i,h}^V$ is conditionally $\bar{\sigma}^2$-sub-Gaussian.

Since $g_{i,h}^{f,V}$ is $\mathcal{F}_{i,h}^a$-measurable, for any $\alpha > 0$,

$$\mathbb{E}\left[\exp\left(\alpha \xi_{i,h}^V g_{i,h}^{f,V} - \frac{\alpha^2 \bar{\sigma}^2}{2} (g_{i,h}^{f,V})^2\right) \,\middle|\, \mathcal{F}_{i,h}^a\right] \leq 1. \tag{77}$$

Iterating (77) over $i = 1, \ldots, t-1$, applying Markov's inequality, and taking a union bound over $h \in [H]$, $t \in [T]$, $V \in \mathcal{V}_{h+1}$, and $f \in \mathcal{R}_h$, we obtain an event of probability at least $1 - \delta/2$ on which, simultaneously over all these indices,

$$\sum_{i=1}^{t-1} \xi_{i,h}^V g_{i,h}^{f,V} \leq \frac{\alpha \bar{\sigma}^2}{2} \sum_{i=1}^{t-1} (g_{i,h}^{f,V})^2 + \frac{1}{\alpha} \log\left(\frac{4THN_{\mathcal{R}_h} N_{\mathcal{V}_{h+1}}}{\delta}\right). \tag{78}$$

Choosing $\alpha = 1/(4\bar{\sigma}^2)$ proves (74). □

Lemma B.10 (Stagewise misspecified ERM localization)**.** *Fix $t \in [T]$ and $h \in [H]$, and set $V = \widetilde{V}_{t,h+1} \in \mathcal{V}_{h+1}$. Let $f^\dagger_{t,h} := f^{\dagger,V}_h$, and let $\hat{Q}_{t,h}$ be the ERM defined by (71). Then, on $\mathcal{E}^{\mathrm{off}}_{\mathrm{RL}}$, simultaneously for all $t \in [T]$ and $h \in [H]$,*

$$\sum_{i=1}^{t-1}\big(\hat{Q}_{t,h}(z_{i,h}) - f^\dagger_{t,h}(z_{i,h})\big)^2 \le C_{\mathrm{rl}}\left[\bar{\sigma}^2 \log\left(\frac{4THN_{\mathcal{R}_h}N_{\mathcal{V}_{h+1}}}{\delta}\right) + T\zeta^2_{\mathrm{RL}}\right], \tag{79}$$

for a universal constant $C_{\mathrm{rl}} > 0$. Consequently, the choice (65) with c_λ sufficiently large implies

$$\sum_{i=1}^{t-1}\big(\hat{Q}_{t,h}(z_{i,h}) - f^\dagger_{t,h}(z_{i,h})\big)^2 \le \lambda \tag{80}$$

simultaneously for all t, h.

Proof. Fix (t,h), set $V = \widetilde{V}_{t,h+1}$, and abbreviate

$$z_i := z_{i,h}, \qquad y_i := y^{(t)}_{i,h}, \qquad m_i := m_{t,h}(z_i), \qquad \xi_i := y_i - m_i,$$

as well as

$$\Delta_i := \hat{Q}_{t,h}(z_i) - f^\dagger_{t,h}(z_i), \qquad g_i := f^\dagger_{t,h}(z_i) - m_{t,h}(z_i).$$

By (63), with $V = \widetilde{V}_{t,h+1}$, we have $|g_i| \le \zeta_{\mathrm{RL}}$. ERM optimality gives

$$\sum_{i=1}^{t-1}(\hat{Q}_{t,h}(z_i) - y_i)^2 \le \sum_{i=1}^{t-1}(f^\dagger_{t,h}(z_i) - y_i)^2. \tag{81}$$

Since $y_i = m_i + \xi_i$, expanding (81) gives

$$\sum_{i=1}^{t-1}\Delta_i^2 \le 2\sum_{i=1}^{t-1}\xi_i\Delta_i - 2\sum_{i=1}^{t-1}g_i\Delta_i. \tag{82}$$

On $\mathcal{E}^{\mathrm{off}}_{\mathrm{RL}}$, Lemma B.9 with $V = \widetilde{V}_{t,h+1}$ and $f = \hat{Q}_{t,h}$ gives

$$\sum_{i=1}^{t-1}\xi_i\Delta_i \le \frac{1}{8}\sum_{i=1}^{t-1}\Delta_i^2 + 4\bar{\sigma}^2\log\left(\frac{4THN_{\mathcal{R}_h}N_{\mathcal{V}_{h+1}}}{\delta}\right). \tag{83}$$

The misspecification term satisfies

$$2\left|\sum_{i=1}^{t-1}g_i\Delta_i\right| \le \frac{1}{4}\sum_{i=1}^{t-1}\Delta_i^2 + 4\sum_{i=1}^{t-1}g_i^2 \le \frac{1}{4}\sum_{i=1}^{t-1}\Delta_i^2 + 4T\zeta^2_{\mathrm{RL}}. \tag{84}$$

Combining (82), (83), and (84), and rearranging, proves (79). The final statement follows from the definition of λ in (65). □

Lemma B.11 (Stagewise frozen-target confidence in RL)**.** *There exists an event $\widetilde{\mathcal{E}}_h$ for each stage $h \in [H]$ such that*

$$\mathbb{P}\Big(\bigcap_{h=1}^{H}\widetilde{\mathcal{E}}_h\Big) \ge 1 - \delta/2,$$

and on $\widetilde{\mathcal{E}}_h$, simultaneously for all episodes $t \in [T]$ and all (s,a),

$$\big|\hat{Q}_{t,h}(s,a) - f^\dagger_{t,h}(s,a)\big| \le \mathrm{unc}_{t,h}(s,a), \tag{85}$$

where $\mathrm{unc}_{t,h}$ is defined in (62).

Proof. Let $\widetilde{\mathcal{E}}_h := \mathcal{E}^{\mathrm{off}}_{\mathrm{RL}}$; the probability statement follows from Lemma B.9 and the fact that $\mathcal{E}^{\mathrm{off}}_{\mathrm{RL}}$ is already simultaneous over all stages.

Fix (t, h) and work on $\mathcal{E}^{\mathrm{off}}_{\mathrm{RL}}$. By Lemma B.10, we have

$$\sum_{i=1}^{t-1} \big(\hat{Q}_{t,h}(z_{i,h}) - f^{\dagger}_{t,h}(z_{i,h})\big)^2 \le \lambda.$$

Hence the pair $(\hat{Q}_{t,h}, f^{\dagger}_{t,h})$ is admissible in the supremum defining $\mathcal{U}_{\mathcal{R}_h}(\lambda; (s,a) \mid \bar{D}^{\mathrm{RL}}_{h,t-1})$, and (85) follows from (62). □

Lemma B.12 (Confidence closure, optimism, and Bellman-residual control)**.** *There exists an event $\mathcal{E}_4$ with $\mathbb{P}(\mathcal{E}_4) \ge 1 - \delta/2$ such that, on $\mathcal{E}_4$, simultaneously for all (t, h) and all (s, a),*

$$\big|\hat{Q}_{t,h}(s,a) - f^{\dagger}_{t,h}(s,a)\big| \le \mathrm{unc}_{t,h}(s,a), \tag{86}$$

$$\big|f^{\dagger}_{t,h}(s,a) - m_{t,h}(s,a)\big| \le \zeta_{\mathrm{RL}}, \tag{87}$$

$$\widetilde{Q}_{t,h}(s,a) - m_{t,h}(s,a) \in [0,\, 2b_{t,h}(s,a)]. \tag{88}$$

Moreover, the optimistic recursion holds:

$$\widetilde{Q}_{t,h}(s,a) \ge Q^{\star}_h(s,a) \qquad \textit{and} \qquad \widetilde{V}_{t,h}(s) \ge V_h(Q^{\star}_h; s) \tag{89}$$

for all t, h and all (s, a).

In particular, defining the on-trajectory Bellman residual

$$e_{t,h} := \widetilde{Q}_{t,h}(s_{t,h}, a_{t,h}) - m_{t,h}(s_{t,h}, a_{t,h}), \tag{90}$$

we have $e^2_{t,h} \le 4b_{t,h}(s_{t,h}, a_{t,h})^2$ almost surely.

Proof. Let

$$\mathcal{E}_4 := \bigcap_{h=1}^{H} \widetilde{\mathcal{E}}_h, \tag{91}$$

where $\widetilde{\mathcal{E}}_h$ is from Lemma B.11. By the probability statement in Lemma B.11, $\mathbb{P}(\mathcal{E}_4) \ge 1 - \delta/2$.

Fix (t, h, s, a) and work on $\mathcal{E}_4$. Equation (86) follows from Lemma B.11, while (87) follows from (63) with $V = \widetilde{V}_{t,h+1}$.

By triangle inequality and $\beta \ge 1$,

$$\big|\hat{Q}_{t,h}(s,a) - m_{t,h}(s,a)\big| \le \mathrm{unc}_{t,h}(s,a) + \zeta_{\mathrm{RL}} \le \beta\, \mathrm{unc}_{t,h}(s,a) + \zeta_{\mathrm{RL}}. \tag{92}$$

Also, by the bounded-range assumption on $\mathcal{R}_h$, $\hat{Q}_{t,h} \in [0, B_h]$. Moreover $m_{t,h}(s,a) \in [0, B_h]$: indeed $\widetilde{Q}_{t,h+1} \in [0, B_{h+1}]$, hence $\widetilde{V}_{t,h+1} \in [0, B_{h+1}]$ by Lemma B.8, and therefore $m_{t,h} = r^{\star}_h + \mathbb{E}[\widetilde{V}_{t,h+1}] \in [0, B_h]$. Thus $|\hat{Q}_{t,h}(s,a) - m_{t,h}(s,a)| \le B_h$. Combining this with (92) and (66) gives

$$\big|\hat{Q}_{t,h}(s,a) - m_{t,h}(s,a)\big| \le b_{t,h}(s,a). \tag{93}$$

Since $\hat{Q}_{t,h}(s,a) + b_{t,h}(s,a) \ge m_{t,h}(s,a)$, $m_{t,h}(s,a) \in [0, B_h]$, and clipping onto $[0, B_h]$ is monotone,

$$\widetilde{Q}_{t,h}(s,a) = \Pi[0, B_h](\hat{Q}_{t,h}(s,a) + b_{t,h}(s,a)) \ge \Pi[0, B_h](m_{t,h}(s,a)) = m_{t,h}(s,a).$$

Also, using non-expansiveness of clipping and (93),

$$\begin{aligned}\widetilde{Q}_{t,h}(s,a) - m_{t,h}(s,a) &\le \left|\hat{Q}_{t,h}(s,a) + b_{t,h}(s,a) - m_{t,h}(s,a)\right| \\ &\le \left|\hat{Q}_{t,h}(s,a) - m_{t,h}(s,a)\right| + b_{t,h}(s,a) \\ &\le 2b_{t,h}(s,a),\end{aligned}$$

which proves (88).

We now prove optimism (89) by backward induction on h. For $h = H$, $m_{t,H}(s,a) = r_H^\star(s,a) = Q_H^\star(s,a)$ since $\widetilde{Q}_{t,H+1} \equiv 0$, and the already proved $\widetilde{Q}_{t,H} \ge m_{t,H}$ gives $\widetilde{Q}_{t,H} \ge Q_H^\star$.

Assume $\widetilde{Q}_{t,h+1} \ge Q_{h+1}^\star$. Monotonicity of $V_{h+1}(\cdot; s)$ yields

$$\widetilde{V}_{t,h+1}(s) = V_{h+1}(\widetilde{Q}_{t,h+1}; s) \ge V_{h+1}(Q_{h+1}^\star; s).$$

Hence

$$m_{t,h} = \mathcal{T}_{\eta,h}\widetilde{Q}_{t,h+1} \ge \mathcal{T}_{\eta,h}Q_{h+1}^\star = Q_h^\star.$$

Since $\widetilde{Q}_{t,h} \ge m_{t,h}$, we obtain $\widetilde{Q}_{t,h} \ge Q_h^\star$. Applying $V_h(\cdot; s)$ yields $\widetilde{V}_{t,h}(s) \ge V_h(Q_h^\star; s)$.

Evaluating (88) at $(s_{t,h}, a_{t,h})$ gives $e_{t,h}^2 \le 4b_{t,h}(s_{t,h}, a_{t,h})^2$. □

Lemma B.13 (Optimism implies quadratic self-bounding of stagewise Gibbs KL)**.** *Fix a stage h and a state s. Let $u(\cdot), v(\cdot)$ be scores on $\mathcal{A}(s)$ such that $u(a) \ge v(a)$ for all $a \in \mathcal{A}(s)$. Let π_u, π_v be the induced Gibbs policies w.r.t. $\pi_{\mathrm{ref},h}(\cdot \mid s)$. Then for any $\eta > 0$,*

$$\frac{1}{\eta}\mathrm{KL}\big(\pi_u(\cdot \mid s) \,\|\, \pi_v(\cdot \mid s)\big) \le \eta\, \mathbb{E}_{a\sim\pi_u(\cdot|s)}\big[(u(a) - v(a))^2\big]. \tag{94}$$

Proof. Let $F(w) := \log \mathbb{E}_{a\sim\pi_{\mathrm{ref},h}(\cdot|s)}[e^{\eta w(a)}]$ and $\Delta := u - v \ge 0$. As before, $\mathrm{KL}(\pi_u\|\pi_v)$ is the Bregman divergence of F, and with $w_\lambda := v + \lambda\Delta$,

$$\mathrm{KL}(\pi_u\|\pi_v) \le \eta^2 \int_0^1 (1-\lambda)\mathbb{E}_{\pi_{w_\lambda}}[\Delta^2]\, d\lambda.$$

Define $\phi(\lambda) := \mathbb{E}_{\pi_{w_\lambda}}[\Delta^2]$. Differentiating Gibbs expectations yields

$$\phi'(\lambda) = \eta\ \mathrm{Cov}_{a\sim\pi_{w_\lambda}}(\Delta(a)^2, \Delta(a)) \ge 0,$$

since Δ^2 is nondecreasing in Δ on $[0,\infty)$. Hence ϕ is nondecreasing and $\mathbb{E}_{\pi_{w_\lambda}}[\Delta^2] \le \mathbb{E}_{\pi_u}[\Delta^2]$ for all $\lambda \in [0,1]$. Therefore,

$$\mathrm{KL}(\pi_u\|\pi_v) \le \eta^2\Big(\int_0^1 (1-\lambda)\, d\lambda\Big)\mathbb{E}_{\pi_u}[\Delta^2] = \frac{\eta^2}{2}\mathbb{E}_{\pi_u}[\Delta^2] \le \eta^2\mathbb{E}_{\pi_u}[\Delta^2].$$

Dividing by η yields (94). □

Lemma B.14 (Deterministic per-episode bound on realized KL regret)**.** *Fix stage h, state s, and two scores $u, v \in [0, B_h]^{\mathcal{A}}$. Let π_u and π_v be the Gibbs policies induced by u and v with respect to $\pi_{\mathrm{ref},h}(\cdot \mid s)$. Then*

$$\frac{1}{\eta}\mathrm{KL}\big(\pi_u(\cdot \mid s) \,\|\, \pi_v(\cdot \mid s)\big) \le 2B_h. \tag{95}$$

Consequently, the per-episode KL regret satisfies $\mathrm{Reg}_\eta^{\mathrm{RL}}(t) \le H(H+1) \le 2H^2$ deterministically.

Proof. Fix s and abbreviate $V_h(\cdot; s)$ by $V(\cdot)$. For Gibbs policies induced by u, v, the KL-gap identity (Lemma B.1 applied pointwise with reference $\pi_{\mathrm{ref},h}(\cdot \mid s)$) implies

$$\frac{1}{\eta}\mathrm{KL}(\pi_u \| \pi_v) = \mathbb{E}_{a\sim\pi_u}[u(a) - v(a)] - \big(V(u) - V(v)\big).$$

Since $u, v \in [0, B_h]$, we have $\mathbb{E}_{\pi_u}[u - v] \le \|u - v\|_\infty \le B_h$, and $V(\cdot)$ is 1-Lipschitz in $\|\cdot\|_\infty$, so $|V(u) - V(v)| \le \|u - v\|_\infty \le B_h$. Therefore $\frac{1}{\eta}\mathrm{KL}(\pi_u \| \pi_v) \le 2B_h$. Summing $2B_h$ over h yields $H(H+1) \le 2H^2$. □

Lemma B.15 (Stagewise width-sum bound via eluder dimension)**.** *For each stage $h \in [H]$, let*

$$u_{t,h} := \mathcal{U}_{\mathcal{R}_h}\big(\lambda; z_{t,h} \mid \bar{D}^{\mathrm{RL}}_{h,t-1}\big).$$

Then deterministically,

$$\sum_{t=1}^{T} u_{t,h}^2 \le c_{\mathrm{w,rl}}\, d(\mathcal{R}_h, \lambda, T), \tag{96}$$

for a universal constant $c_{\mathrm{w,rl}} > 0$.

Proof. This is the stagewise counterpart of Lemma B.6. The proof is the same dyadic peeling + eluder counting argument applied to the stage-h trajectory $\{z_{t,h}\}_{t=1}^{T}$ and the class $\mathcal{R}_h$, using the width definition (61). □

Lemma B.16 (Global uncertainty-sum via the width-based RL definition)**.** *Let $z_{t,h} = (s_{t,h}, a_{t,h})$, and define*

$$d_{\mathrm{RL}}(\lambda, T) := \sum_{h=1}^{H} d(\mathcal{R}_h, \lambda, T). \tag{97}$$

Then deterministically,

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \mathrm{unc}_{t,h}(z_{t,h})^2 = \sum_{h=1}^{H}\sum_{t=1}^{T} u_{t,h}^2 \le c_{\mathrm{unc}}\, d_{\mathrm{RL}}(\lambda, T), \tag{98}$$

for a universal constant $c_{\mathrm{unc}} > 0$. Consequently, for the bonus (66),

$$\sum_{t=1}^{T}\sum_{h=1}^{H} b_{t,h}(z_{t,h})^2 \le c_w\Big(\beta^2 d_{\mathrm{RL}}(\lambda, T) + H\,T\zeta^2_{\mathrm{RL}}\Big), \tag{99}$$

for a universal constant $c_w > 0$.

Proof. By (62), $\mathrm{unc}_{t,h}(z_{t,h}) = u_{t,h}$, so

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \mathrm{unc}_{t,h}(z_{t,h})^2 = \sum_{h=1}^{H}\sum_{t=1}^{T} u_{t,h}^2.$$

Applying Lemma B.15 stagewise and summing over h gives (98).

For the bonus bound, by (66) and $(u+v)^2 \le 2u^2 + 2v^2$,

$$b_{t,h}(z_{t,h})^2 \le 2\beta^2\, \mathrm{unc}_{t,h}(z_{t,h})^2 + 2\zeta^2_{\mathrm{RL}}.$$

Summing over (t, h) and applying (98) yields (99) after absorbing constants. □

Lemma B.17 (KL-RL bridge: from squared Q-gaps to squared Bellman residuals)**.** *Fix an episode t and work on $\mathcal{E}_4$ from Lemma B.12. Define the on-trajectory Bellman residuals*

$$e_{t,h} := \widetilde{Q}_{t,h}(s_{t,h}, a_{t,h}) - m_{t,h}(s_{t,h}, a_{t,h}), \tag{100}$$

and the stagewise squared Q-gaps under $\pi_{t,h}$,

$$\Delta_{t,h} := \mathbb{E}_{a\sim\pi_{t,h}(\cdot|s_{t,h})}\big(\widetilde{Q}_{t,h}(s_{t,h}, a) - Q^\star_h(s_{t,h}, a)\big)^2. \tag{101}$$

Then, conditioning on $\mathcal{F}^-_{t,1}$,

$$\sum_{h=1}^{H} \mathbb{E}\big[\Delta_{t,h} \mid \mathcal{F}^-_{t,1}\big] \le H^2 \sum_{h=1}^{H} \mathbb{E}\big[e^2_{t,h} \mid \mathcal{F}^-_{t,1}\big]. \tag{102}$$

Proof. This is the KL-RL extension of the policy-switch bridge used in prior KL-regularized RL regret analyses, but stated directly in terms of the algorithmic optimistic scores $\widetilde{Q}_{t,h}$ and the KL Bellman targets $m_{t,h}$.

Fix episode t, and abbreviate

$$s_h := s_{t,h}, \quad a_h := a_{t,h}, \quad \mathcal{F}^-_h := \mathcal{F}^-_{t,h}, \quad \widetilde{Q}_h := \widetilde{Q}_{t,h}, \quad \widetilde{V}_h := \widetilde{V}_{t,h}, \quad m_h := m_{t,h}, \quad \pi_h := \pi_{t,h}.$$

Define the nonnegative Q-gap and value gap

$$g_h(s,a) := \widetilde{Q}_h(s,a) - Q^\star_h(s,a) \ge 0, \qquad \delta_h(s) := \widetilde{V}_h(s) - V_h(Q^\star_h; s) \ge 0,$$

where nonnegativity follows from Lemma B.12. Let $\delta_{H+1} \equiv 0$.

Since

$$m_h(s,a) = r^\star_h(s,a) + \mathbb{E}[\widetilde{V}_{h+1}(s') \mid s,a], \qquad Q^\star_h(s,a) = r^\star_h(s,a) + \mathbb{E}[V_{h+1}(Q^\star_{h+1}; s') \mid s,a],$$

we have

$$g_h(s,a) = \underbrace{\widetilde{Q}_h(s,a) - m_h(s,a)}_{=:e_h(s,a)} + \mathbb{E}\big[\delta_{h+1}(s') \mid s,a\big]. \tag{103}$$

Moreover, by convexity of $u \mapsto V_h(u; s)$ and the fact that $\pi_h(\cdot \mid s)$ is the Gibbs distribution induced by $\widetilde{Q}_h(s,\cdot)$, the supporting-hyperplane inequality yields

$$\delta_h(s) \le \mathbb{E}_{a\sim\pi_h(\cdot|s)}\big[g_h(s,a)\big]. \tag{104}$$

Step 1: pathwise domination by future Bellman residuals. Fix h, condition on $\mathcal{F}^-_h$, and draw $a_h \sim \pi_h(\cdot \mid s_h)$. Let $(s_j, a_j)_{j>h}$ be the future trajectory generated by $\{\pi_j\}_{j>h}$ and the MDP dynamics. We claim that on $\mathcal{E}_4$,

$$g_h(s_h, a_h) \le \mathbb{E}\left[\sum_{j=h}^{H} e_j(s_j, a_j) \,\middle|\, s_h, a_h\right]. \tag{105}$$

The proof is by backward induction. For $h = H$, $\delta_{H+1} \equiv 0$, so (103) gives $g_H(s_H, a_H) = e_H(s_H, a_H)$. For the induction step, using (103) and (104),

$$\begin{aligned} g_h(s_h, a_h) &= e_h(s_h, a_h) + \mathbb{E}[\delta_{h+1}(s_{h+1}) \mid s_h, a_h] \\ &\le e_h(s_h, a_h) + \mathbb{E}\Big[\mathbb{E}_{a\sim\pi_{h+1}(\cdot|s_{h+1})}[g_{h+1}(s_{h+1}, a)] \,\Big|\, s_h, a_h\Big] \\ &= e_h(s_h, a_h) + \mathbb{E}\big[g_{h+1}(s_{h+1}, a_{h+1}) \mid s_h, a_h\big], \end{aligned}$$

and the induction hypothesis closes the recursion.

Step 2: square and aggregate. By Jensen and Cauchy–Schwarz,

$$g_h(s_h,a_h)^2 \le \mathbb{E}\left[\Big(\sum_{j=h}^{H} e_j(s_j,a_j)\Big)^2 \;\Big|\; s_h,a_h\right] \le (H-h+1)\,\mathbb{E}\left[\sum_{j=h}^{H} e_j(s_j,a_j)^2 \;\Big|\; s_h,a_h\right].$$

Taking expectation over $a_h \sim \pi_h(\cdot \mid s_h)$ conditional on $\mathcal{F}_h^-$ yields

$$\Delta_h \le (H-h+1)\,\mathbb{E}\left[\sum_{j=h}^{H} e_j(s_j,a_j)^2 \;\Big|\; \mathcal{F}_h^-\right].$$

Finally, take $\mathcal{F}_1^-$-conditional expectation and sum over h:

$$\sum_{h=1}^{H}\mathbb{E}[\Delta_h \mid \mathcal{F}_1^-] \le \sum_{j=1}^{H}\Big(\sum_{h=1}^{j}(H-h+1)\Big)\mathbb{E}[e_j^2 \mid \mathcal{F}_1^-] \le H^2\sum_{h=1}^{H}\mathbb{E}[e_h^2 \mid \mathcal{F}_1^-],$$

which proves (102). □

Lemma B.18 (KL-regularized performance-difference identity)**.** *Fix any nonstationary policy $\pi = \{\pi_h\}_{h=1}^H$. Let $\pi^\star$ be the Gibbs policy induced by $Q^\star$. For any initial state s_1,*

$$V_1^\star(s_1) - V_1^\pi(s_1) = \mathbb{E}_\pi\left[\sum_{h=1}^{H}\frac{1}{\eta}\mathrm{KL}(\pi_h(\cdot \mid s_h)\,\|\,\pi_h^\star(\cdot \mid s_h))\;\Big|\; s_1\right]. \tag{106}$$

Proof. For each stage h, the soft Bellman optimality equation and Lemma B.1 applied pointwise with reference $\pi_{\mathrm{ref},h}(\cdot \mid s)$ imply

$$V_h^\star(s) = \mathbb{E}_{a\sim\pi_h(\cdot|s)}\left[Q_h^\star(s,a) - \frac{1}{\eta}\log\frac{\pi_h(a \mid s)}{\pi_{\mathrm{ref},h}(a \mid s)}\right] + \frac{1}{\eta}\mathrm{KL}(\pi_h(\cdot \mid s)\,\|\,\pi_h^\star(\cdot \mid s)). \tag{107}$$

Using $Q_h^\star(s,a) = r_h^\star(s,a) + \mathbb{E}[V_{h+1}^\star(s_{h+1}) \mid s,a]$, subtracting the policy-evaluation recursion for V_h^π, and taking expectation under π, the value-difference terms telescope from $h=1$ to H. This gives (106). □

Lemma B.19 (Conditional-regret reduction to conditional bonus squares)**.** *Define the realized and conditional per-episode KL regrets*

$$R_t := \mathrm{Reg}_\eta^{\mathrm{RL}}(t), \qquad \bar{R}_t := \mathbb{E}[R_t \mid \mathcal{F}_{t,1}^-],$$

and define

$$X_{t,h} := b_{t,h}(s_{t,h},a_{t,h})^2, \qquad \bar{X}_{t,h} := \mathbb{E}[X_{t,h} \mid \mathcal{F}_{t,h}^-] = \mathbb{E}_{a\sim\pi_{t,h}(\cdot|s_{t,h})}\big[b_{t,h}(s_{t,h},a)^2\big]. \tag{108}$$

Then on $\mathcal{E}_4$,

$$\sum_{t=1}^{T}\bar{R}_t \le 4\eta H^2\sum_{t=1}^{T}\sum_{h=1}^{H}\bar{X}_{t,h}. \tag{109}$$

Proof. Fix t. By Lemma B.18 applied to $\pi = \pi_t$, the conditional regret $\bar{R}_t$ is the $\mathcal{F}_{t,1}^-$-conditional expectation of the sum of stagewise KL gaps against $\pi^\star$.

Fix (t,h). Conditioning on $s_{t,h}$, Lemma B.12 gives $\widetilde{Q}_{t,h}(s_{t,h},\cdot) \geq Q_h^\star(s_{t,h},\cdot)$. Applying Lemma B.13,

$$\frac{1}{\eta}\mathrm{KL}\big(\pi_{t,h}(\cdot \mid s_{t,h}) \| \pi_h^\star(\cdot \mid s_{t,h})\big) \leq \eta\, \Delta_{t,h}.$$

Summing over h and taking $\mathcal{F}_{t,1}^-$-conditional expectation yields

$$\bar{R}_t \leq \eta \sum_{h=1}^{H} \mathbb{E}\big[\Delta_{t,h} \mid \mathcal{F}_{t,1}^-\big].$$

By Lemma B.17,

$$\sum_{h=1}^{H} \mathbb{E}[\Delta_{t,h} \mid \mathcal{F}_{t,1}^-] \leq H^2 \sum_{h=1}^{H} \mathbb{E}[e_{t,h}^2 \mid \mathcal{F}_{t,1}^-].$$

On $\mathcal{E}_4$, Lemma B.12 gives $e_{t,h}^2 \leq 4X_{t,h}$. By tower property,

$$\mathbb{E}[e_{t,h}^2 \mid \mathcal{F}_{t,1}^-] \leq 4\,\mathbb{E}[X_{t,h} \mid \mathcal{F}_{t,1}^-] = 4\,\mathbb{E}[\bar{X}_{t,h} \mid \mathcal{F}_{t,1}^-].$$

Combining the displays and summing over t yields (109). □

Lemma B.20 (Stage-action bonus-square alignment (Freedman))**.** *With $X_{t,h}, \bar{X}_{t,h}$ as defined in (108), there exists an event with probability at least $1-\delta/4$ on which*

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \bar{X}_{t,h} \leq 2\sum_{t=1}^{T}\sum_{h=1}^{H} X_{t,h} + 4H^2 \log\Big(\frac{8}{\delta}\Big). \tag{110}$$

Proof. Index (t,h) lexicographically by $k = (t-1)H + h$, and let $K := TH$. Write $\widetilde{\mathcal{F}}_k := \mathcal{F}_{t,h}^-$. Then $X_{t,h}$ is $\widetilde{\mathcal{F}}_{k+1}$-measurable, while $\bar{X}_{t,h}$ is $\widetilde{\mathcal{F}}_k$-measurable.

Define the scaled shifted martingale differences

$$Y_{k+1} := \frac{X_{t,h} - \bar{X}_{t,h}}{H^2} \in [-1,1].$$

Then $\mathbb{E}[Y_{k+1} \mid \widetilde{\mathcal{F}}_k] = 0$. Since $X_{t,h} \in [0, H^2]$, we have $X_{t,h}^2 \leq H^2 X_{t,h}$, and thus

$$\mathbb{E}[Y_{k+1}^2 \mid \widetilde{\mathcal{F}}_k] = \frac{\mathbb{E}[(X_{t,h} - \bar{X}_{t,h})^2 \mid \widetilde{\mathcal{F}}_k]}{H^4} \leq \frac{\mathbb{E}[X_{t,h}^2 \mid \widetilde{\mathcal{F}}_k]}{H^4} \leq \frac{\bar{X}_{t,h}}{H^2}.$$

Let

$$M_{K+1} := \sum_{k=1}^{K} Y_{k+1}, \qquad V := \sum_{k=1}^{K} \mathbb{E}[Y_{k+1}^2 \mid \widetilde{\mathcal{F}}_k] \leq \frac{1}{H^2}\sum_{t=1}^{T}\sum_{h=1}^{H} \bar{X}_{t,h}.$$

Applying Freedman's inequality to $-M_{K+1}$, with failure probability $\delta/4$, gives

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \frac{\bar{X}_{t,h} - X_{t,h}}{H^2} \leq \sqrt{2V\log(8/\delta)} + \frac{1}{3}\log\Big(\frac{8}{\delta}\Big).$$

Using $\sqrt{2V\log(8/\delta)} \leq \frac{1}{2}V + \log(8/\delta)$, multiplying by H^2, and rearranging yields (110) up to slightly looser constants. □

Lemma B.21 (Episode-level realized-vs-conditional regret alignment (Freedman))**.** *Let* $R_t := \mathrm{Reg}_\eta^{\mathrm{RL}}(t)$ *and* $\bar{R}_t := \mathbb{E}[R_t \mid \mathcal{F}_{t,1}^-]$*. There exists an event with probability at least* $1 - \delta/4$ *on which*

$$\sum_{t=1}^{T} R_t \le 2\sum_{t=1}^{T} \bar{R}_t + 8H^2 \log\Big(\frac{8}{\delta}\Big). \tag{111}$$

Proof. By Lemma B.14, $R_t \le 2H^2$ deterministically. Let

$$X_t := \frac{R_t}{2H^2} \in [0,1], \qquad \bar{X}_t := \mathbb{E}[X_t \mid \mathcal{F}_{t,1}^-] = \frac{\bar{R}_t}{2H^2}.$$

As in the bandit case, X_t is $\mathcal{F}_{t+1,1}^-$-measurable and $\bar{X}_t$ is $\mathcal{F}_{t,1}^-$-measurable. Define $Y_{t+1} := X_t - \bar{X}_t$. Then Y_{t+1} is a shifted martingale difference w.r.t. $\{\mathcal{F}_{t,1}^-\}_{t\ge 1}$, with $|Y_{t+1}| \le 1$ and

$$\mathbb{E}[Y_{t+1}^2 \mid \mathcal{F}_{t,1}^-] \le \mathbb{E}[X_t \mid \mathcal{F}_{t,1}^-] = \bar{X}_t.$$

Let

$$M_{T+1} := \sum_{t=1}^{T} Y_{t+1}, \qquad V := \sum_{t=1}^{T} \mathbb{E}[Y_{t+1}^2 \mid \mathcal{F}_{t,1}^-] \le \sum_{t=1}^{T} \bar{X}_t.$$

Freedman's inequality applied to $-M_{T+1}$, with failure probability $\delta/4$, yields

$$\sum_{t=1}^{T} (\bar{X}_t - X_t) \le \sqrt{2V \log(8/\delta)} + \frac{1}{3}\log\Big(\frac{8}{\delta}\Big).$$

Using $\sqrt{2V\log(8/\delta)} \le \frac{1}{2}V + \log(8/\delta)$ and rearranging gives

$$\sum_{t=1}^{T} X_t \le 2\sum_{t=1}^{T} \bar{X}_t + 2\log\Big(\frac{8}{\delta}\Big).$$

Multiplying by $2H^2$ yields (111). □

Proof of Theorem 5.4. Work on the intersection of: (i) $\mathcal{E}_4$ from Lemma B.12, (ii) the stage-action bonus-square alignment event in Lemma B.20, and (iii) the episode-level regret alignment event in Lemma B.21. By Lemma B.12, $\mathbb{P}(\mathcal{E}_4^c) \le \delta/2$, and the other two events each fail with probability at most $\delta/4$. Hence a union bound gives total failure probability at most δ.

By Lemma B.19, on $\mathcal{E}_4$,

$$\sum_{t=1}^{T} \bar{R}_t \le 4\eta H^2 \sum_{t=1}^{T}\sum_{h=1}^{H} \bar{X}_{t,h}.$$

Lemma B.20 yields

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \bar{X}_{t,h} \le 2\sum_{t=1}^{T}\sum_{h=1}^{H} X_{t,h} + 4H^2 \log\Big(\frac{8}{\delta}\Big).$$

Applying Lemma B.16,

$$\sum_{t=1}^{T}\sum_{h=1}^{H} X_{t,h} = \sum_{t=1}^{T}\sum_{h=1}^{H} b_{t,h}(z_{t,h})^2 \le c_w\Big(\beta^2 d_{\mathrm{RL}}(\lambda, T) + H\,T\zeta_{\mathrm{RL}}^2\Big). \tag{112}$$

Therefore, on the intersection event,

$$\sum_{t=1}^{T} \bar{R}_t \le C\,\eta H^2\Big(\beta^2 d_{\mathrm{RL}}(\lambda, T) + H\,T\zeta_{\mathrm{RL}}^2 + H^2 \log\tfrac{1}{\delta}\Big), \tag{113}$$

for a universal constant $C > 0$.

Finally, on the event (111),

$$\sum_{t=1}^{T} R_t \le 2 \sum_{t=1}^{T} \bar{R}_t + 8H^2 \log\Big(\frac{8}{\delta}\Big),$$

which together with (113) yields (31), up to the lower-order logarithmic terms hidden in the theorem statement. □