PROVABLE BENEFITS OF REGULARIZATION: FAST RATES FOR ADVERSARIAL IMITATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We study adversarial imitation learning (AIL), in which an agent learns to imitate expert demonstrations by optimizing a policy against an adversarial reward that distinguishes expert and learner behavior. Historically, reward regularization and entropy-based policy regularization are key components of empirically successful methods such as GAIL and LS-IQ, yet their finite-sample benefits remain underexplored. We establish fast rates for jointly regularized AIL in finite-horizon Markov decision processes with general function approximation. Our model-free algorithm, Dually Regularized AIL, combines KL policy regularization with a quadratic reward penalty weighted by expert and learner occupancies. With K online episodes and N expert trajectories, we prove a $\widetilde{\mathcal{O}}\left(\frac{1}{K}+\frac{1}{N}\right)$ bound on the regularized imitation gap for fixed regularization parameters. Our analysis combines an online mirror descent construction for general convex reward classes to control estimation error from finite expert data and stochastic learner feedback, with a sharp analysis of optimistic KL-regularized policy learning. To the best of our knowledge, Dually Regularized AIL is the first algorithm to simultaneously achieve $\widetilde{\mathcal{O}}\left(\frac{1}{\epsilon}\right)$ sample complexity in both expert demonstrations and online interactions for this regularized AIL objective, even with stochastic experts. These results provide a rigorous characterization of the complementary statistical benefits of reward and policy regularization in AIL.

1 INTRODUCTION

Imitation learning (IL) seeks to learn a policy that reproduces expert behavior from demonstrations, with the goal of matching the expert's performance when the task reward is unavailable. Behavior cloning approaches this goal by directly predicting expert actions. Adversarial imitation learning (AIL) instead learns a reward that distinguishes expert and learner behavior, and improves the policy under that reward (Abbeel & Ng, 2004; Syed & Schapire, 2007; Ho & Ermon, 2016). This provides a performance-based criterion for imitation: expected cumulative reward depends on the state–action occupancy measure, so reducing the expert–learner gap over a reward class controls performance on the tasks represented by that class. Reward learning therefore guides the policy toward matching the expert's behavior across the trajectory.

As AIL has developed, reward and policy regularization have become important ingredients of practical algorithms. A common motivation for reward regularization is to reduce overfitting to finite demonstrations and maintain a useful learning signal for the policy (Orsini et al., 2021). Entropy-based policy regularization complements this by encouraging stochastic behavior and exploration during policy improvement (Haarnoja et al., 2018). GAIL incorporates both ingredients through cost-regularized occupancy matching and a causal-entropy policy objective (Ho & Ermon, 2016). Related inverse-Q methods retain this two-sided structure: IQ-Learn combines implicit reward regularization with soft-Q or soft actor–critic policy updates, while LS-IQ studies a quadratic reward penalty under a mixture of expert and learner occupancies within a maximum-entropy formulation (Garg et al., 2021; Al-Hafez et al., 2023). These developments motivate regularization as a way to limit overfitting and stabilize learning, and connect its design to the underlying occupancy-matching objective. How these benefits translate into faster convergence from finite demonstrations and online interactions remains less understood.

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

One way to make these statistical benefits precise is to examine whether regularization can improve convergence rates. Recent advances in IL and reinforcement learning (RL) show that structural assumptions and regularized objectives can yield faster rates than the conventional inverse-square-root dependence on sample size. For realizable deterministic experts, Foster et al. (2024) establish a $\widetilde{\mathcal{O}}(1/N)$ performance bound for log-loss behavior cloning from N expert trajectories. In online RL, Zhao et al. (2025) exploit the curvature of a KL-regularized policy objective to obtain logarithmic cumulative regret and a corresponding $\widetilde{\mathcal{O}}(1/K)$ average policy error after K episodes. These results motivate examining whether reward regularization can similarly accelerate the convergence of AIL.

However, one challenging part in the AIL is that the coupling between the reward and the policy. In particular, since the reward is learned from the same finite expert dataset throughout training, while the evolving policy determines the distribution of newly collected learner trajectories. The reward learner must therefore control estimation errors from both data sources, and the policy learner must improve its behavior using an estimated value function. This coupling brings the two regularizers into the same statistical problem and raises the following question:

Can reward and policy regularization jointly yield fast rates for AIL in both expert demonstrations and online interactions?

In this paper, we answer this question for jointly regularized AIL in finite-horizon MDPs with unknown transitions and general function approximation. We propose Dually Regularized AIL, a model-free algorithm combining KL policy regularization with a quadratic reward penalty weighted by expert and learner occupancies. The policy objective recovers causal-entropy regularization under a uniform reference. The algorithm alternates optimistic value-based policy learning with online mirror descent over a general convex reward class, reusing a fixed expert dataset and collecting one new learner trajectory per episode. We summarize our technical contributions as follows:

Fast rates for regularized AIL. With K online episodes and N expert trajectories, Theorem 1 and Corollary 1 establish the high-probability bound $\mathrm{Gap}(\bar{\pi}_K, \bar{r}_K) \le \widetilde{\mathcal{O}}\left(\frac{1}{K} + \frac{1}{N}\right)$ on the regularized saddle-point gap, for fixed positive regularization parameters and controlled function-class complexity. This yields $\widetilde{\mathcal{O}}(\epsilon^{-1})$ sample requirements for both expert demonstrations and online interactions, even for stochastic experts. The full theorem specifies the dependence on the horizon, regularization strengths, covering numbers, and generalized eluder dimensions.

Statistical roles of regularization. Our reward analysis uses an online mirror descent construction whose update stability is controlled by generalized eluder dimension. The occupancy-weighted quadratic penalty supplies curvature that controls estimation error from the reused expert dataset and stochastic feedback from learner trajectories. We combine this construction with the sharp optimistic KL-RL analysis of Zhao et al. (2025), in which policy curvature converts planning uncertainty into squared Bellman errors. Together, these arguments explain how reward and policy regularization contribute to fast finite-sample rates in AIL.

2 Related Work

In this section, we provide an overview of the related works in AIL and RL settings. Table 1 summarizes representative guarantees, separating their objectives and expert assumptions.

Reward regularization and inverse Q-learning. Apprenticeship learning connects reward learning to expert-performance matching (Abbeel & Ng, 2004; Syed & Schapire, 2007). GAIL formalizes the connection between a convex cost regularizer and the induced occupancy discrepancy through convex conjugacy; its particular choice gives a Jensen–Shannon objective together with causal-entropy regularization (Ho & Ermon, 2016). IQ-Learn represents reward and policy through a single soft Q-function, avoiding explicit alternation between a separately parameterized reward and policy (Garg et al., 2021) which can be considered as the max-entropy RL for policy regularization. Building on this perspective, LS-IQ studies a squared reward penalty under the expert–learner mixture and relates it to a bounded Pearson χ^2 divergence, reward bounds, and improved stability (Al-Hafez et al., 2023). Empirical comparisons also show that discriminator regularization can matter substantially, especially on harder tasks, although its effect depends on other algorithmic choices (Orsini et al., 2021). These works motivate our regularizer but address a different question from the joint finite-

Table 1: Sample complexity in expert trajectories (N) and online episodes (K) in related works. Sample complexity omits the horizon, reward-range, and structural-complexity factors. "General" experts may be stochastic. All guarantees are high-probability except Mimic-Emp's expected gap. Unregularized rates summarize the worst case, omitting variance-dependent improvements. These works made regular RL / IL assumptions where we refer readers to check their original statements.

Method	Function class	Expert	Criterion	N	K
Mimic-Emp (Rajaraman et al., 2020)	Tabular	General	Expected IL gap	ϵ^{-1}	0
Log-loss BC (Foster et al., 2024)	General	Deterministic	Imitation gap	ϵ^{-1}	0
Log-loss BC (Foster et al., 2024)	General	General	Imitation gap	ϵ^{-2}	0
MB-TAIL (Xu et al., 2023)	Tabular	Deterministic	Imitation gap	ϵ^{-1}	ϵ^{-2}
OPT-AIL (Xu et al., 2024)	General	General	Imitation gap	ϵ^{-2}	ϵ^{-2}
MB-AIL (Li et al., 2026)	General	General	Imitation gap	ϵ^{-2}	ϵ^{-2}
KL-LSVI-UCB (Zhao et al., 2025)	General	Not needed	Regularized RL regret	–	ϵ^{-1}
Dually Regularized AIL (ours)	**General**	**General**	**Reg. AIL dual gap**	$\boldsymbol{\epsilon^{-1}}$	$\boldsymbol{\epsilon^{-1}}$

sample rates proved here. In particular, a quadratic penalty controls reward magnitude and supplies curvature; it need not enforce smoothness with respect to state–action inputs.

Finite-sample adversarial imitation learning. Online apprenticeship learning combines optimistic exploration with no-regret reward and policy updates, yielding square-root interaction regret and an additional error from finite expert data (Shani et al., 2022). Subsequent analyses cover unknown tabular dynamics (Xu et al., 2023), linear-mixture models (Liu et al., 2021), and linear MDPs (Viano et al., 2024). Under general function approximation, OPT-AIL couples reward optimization with optimistic policy learning and characterizes complexity using a generalized eluder coefficient (Xu et al., 2024). MB-AIL instead learns a transition model and establishes second-order guarantees that adapt to stochasticity, together with lower bounds on expert and interaction requirements (Li et al., 2026). Its worst-case square-root terms can improve in favorable low-variance instances. Our result studies a regularized saddle-point criterion and uses reward curvature to obtain inverse-sample dependence without requiring a deterministic expert or vanishing variance. The different objectives and structural assumptions are essential to interpreting the comparison.

Fast rates and the role of expert stochasticity. Fast expert-sample rates depend on both the policy class and the expert's stochasticity. In tabular MDPs, Rajaraman et al. (2020) obtain inverse-N expected-error bounds even for stochastic experts. For general policy classes, the worst-case picture is different. For a realizable deterministic expert, Foster et al. (2024) combine log-loss estimation with a trajectory-level analysis to obtain $\tilde{\mathcal{O}}(N^{-1})$ performance error; stochastic experts admit variance-sensitive bounds with a worst-case $N^{-1/2}$ term. Their deterministic-expert result does not require deterministic transitions. Fast expert-sample rates also arise in model-based tabular AIL with deterministic experts (Xu et al., 2023), while the second-order AIL guarantees of Li et al. (2026) quantify more general instance-dependent improvements. Our fast rate arises from the regularized objective rather than expert determinism. It therefore complements these results instead of contradicting lower bounds for ordinary imitation. In particular, the success of offline cloning rules out a universal claim that online interaction is necessary; our result characterizes the efficiency of a reward-based alternative.

3 Preliminaries

We consider an episodic Markov decision process (MDP) $\langle \mathcal{S}, \mathcal{A}, H, P \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ are the state space and action space, $H \in \mathbb{Z}_+$ is the horizon, and $P = \{P_h\}_{h\in[H]}$ with $P_h : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ is the transition dynamics. We assume the initial state is fixed from s_1. A policy is denoted by $\pi = \{\pi_h\}_{h\in[H]}$, where $\pi_h : \mathcal{S} \to \Delta_{\mathcal{A}}$. For each $h \in [H]$, let d_h^π denote the state-action occupancy distribution of a state-action pair (s_h, a_h) induced by P and π at timestep h. We write $\langle \mu, f \rangle := \int f \mathrm{d}\mu$ for a signed measure μ and a bounded function f as the expectation $\mathbb{E}_\mu[f(\cdot)]$ for simplicity.

In **imitation learning**, consider π^E be the expert policy and denote its occupancy by d_h^E. We observe an expert dataset $\mathcal{D}^E = \{s_h^i, a_h^i\}_{h\in[H]}^{i\in[N]}$ containing N independent expert trajectories. The agent then performs K-round of online interactions with the environments. In the AIL setting, the agent seeks a stage-wise reward function $r = \{r_h\}_{h\in[H]}$ in order to distinguish the expert demonstrations and the learner behavior. We assume the reward function in this class bounded by $r(s,a) \in [-1,1]$.

KL regularized policy updates. Let π^{ref} be a reference policy that is known to the learner and let $\tau > 0$ be the KL-regularization coefficient. The reference policy π^{ref} may be initialized from a policy pretrained via behavioral cloning, similar to prior empirical approaches (Jena et al., 2021). Alternatively, choosing π^{ref} to be uniform reduces KL regularization to entropy regularization up to an additive constant, recovering the maximum-entropy AIL formulation commonly used in prior work, including GAIL (Ho & Ermon, 2016) and IQ-Learn (Garg et al., 2021). We use the convention $0\log(0/q) = 0$, with KL divergence equal to $+\infty$ when its first argument assigns positive mass outside the support of its second. Gibbs identities are understood on the support of π^{ref}. Following prior work (Zhao et al., 2025), define the KL-regularized value and action-value functions for a policy π under any bounded stagewise reward r as

$$V_{r,h}^{\pi,\tau}(s) := \mathbb{E}_{P,\pi}\left[\sum_{t=h}^{H}\left(r_t(s_t,a_t) - \tau\log\pi_t(a_t \mid s_t)/\pi_t^{\mathrm{ref}}(a_t \mid s_t)\right)\,\middle|\, s_h = s\right],$$
$$Q_{r,h}^{\pi,\tau}(s,a) := r_h(s,a) + [P_h V_{r,h+1}^{\pi,\tau}](s,a),$$

where $V_{r,H+1}^{\pi,\tau} \equiv 0$ and $[P_hV](s,a) := \mathbb{E}_{s'\sim P_h(\cdot|s,a)}[V(s')]$. Equivalently,

$$V_{r,h}^{\pi,\tau}(s) = \mathbb{E}_{a\sim\pi_h(\cdot|s)}\left[Q_{r,h}^{\pi,\tau}(s,a)\right] - \tau\,\mathrm{KL}\big(\pi_h(\cdot \mid s)\,\|\,\pi_h^{\mathrm{ref}}(\cdot \mid s)\big).$$

Regularized reward update While KL-divergence to a reference policy characterizes the policy-side regularization, for a fixed reward, our setup focuses on adversarial imitation learning (AIL), in which the objective compares the learner with the expert uniformly over the reward class $\mathcal{R}$. We complement the policy-side KL regularization with a quadratic reward regularizer, following prior practical AIL methods such as LS-IQ (Al-Hafez et al., 2023). For an expert occupancy weight $\omega \in (0,1)$, define the quadratic reward regularizer

$$\psi_\pi^\omega(r) := \frac{\alpha}{2}\sum_{h=1}^{H}\left[\omega\mathbb{E}_{d_h^E}[r_h^2] + (1-\omega)\mathbb{E}_{d_h^\pi}[r_h^2]\right], \qquad \alpha > 0.$$

For any reward r, write $J_r^\tau(\pi) := V_{r,1}^{\pi,\tau}(s_1)$. The regularized AIL objective can be written by the following min-max optimization on policy π and reward r:

$$\max_{r\in\mathcal{R}}\min_\pi \mathcal{L}(\pi,r) := J_r^\tau(\pi^E) - J_r^\tau(\pi) - \psi_\pi^\omega(r) \tag{1}$$

We then define the regret over K online episodes by the cumulative dual gap as

$$\mathrm{Regret}_{\omega,\alpha,\tau}(K) := \sup_{r\in\mathcal{R}}\sum_{k=1}^{K}\mathcal{L}(\pi_k,r) - \inf_\pi\sum_{k=1}^{K}\mathcal{L}(\pi,r_k). \tag{2}$$

AIL with Function Approximation. We consider model-free adversarial imitation learning under general function approximation. Specifically, let $\mathcal{F} = \{\mathcal{F}_h\}_{h\in[H]}$, where each $\mathcal{F}_h$ is a nonempty class of functions from $\mathcal{S}\times\mathcal{A}$ to $[-2V_{\max}, 2V_{\max}]$ for $V_{\max} := H\left(1 + \frac{\alpha(1-\omega)}{2}\right)$. The algorithm fits its regression functions in these classes. To characterize the exploration complexity of the underlying MDP, we adopt the generalized Eluder dimension introduced by Agarwal et al. (2023):

Definition 1 (Generalized eluder dimension). *Let $\lambda > 0$, $h \in [H]$, and $Z_h = (z_{k,h})_{k=1}^K$, where $z_{k,h} = (s_{k,h}, a_{k,h})$, be an indexed sequence of state–action pairs, with repetitions retained. Write $z_{[k-1],h} := (z_{j,h})_{j=1}^{k-1}$. For a bounded class $\mathcal{F}_h$ of real-valued functions on $\mathcal{S}\times\mathcal{A}$, define*

$$\dim_K(\mathcal{F}_h,\lambda) := \sup_{|Z_h|=K}\sum_{k=1}^{K}\min\left\{1, D_{\mathcal{F}_h}^2(z_{k,h}; z_{[k-1],h}; \lambda)\right\},$$
$$D_{\mathcal{F}_h}^2(z_{k,h}; z_{[k-1],h}; \lambda) := \sup_{f_1,f_2\in\mathcal{F}_h}\frac{(f_1(z_{k,h}) - f_2(z_{k,h}))^2}{\sum_{j\in[k-1]}(f_1(z_{j,h}) - f_2(z_{j,h}))^2 + \lambda}.$$

We write $D_{\mathcal{F}_h} := \sqrt{D_{\mathcal{F}_h}^2}$ and denote $\dim_K(\mathcal{F},\lambda) = H^{-1}\sum_{h\in[H]}\dim_K(\mathcal{F}_h,\lambda)$ for simplicity. These definitions also apply to $\mathcal{R}_h$, with $\dim_K(\mathcal{R},\lambda) := H^{-1}\sum_{h=1}^{H}\dim_K(\mathcal{R}_h,\lambda)$.

Algorithm 1 Dually Regularized Adversarial Imitation Learning

1: **Input:** Expert dataset $\mathcal{D}^E$, convex reward class $\mathcal{R}$, Q-function class $\mathcal{F} = \{\mathcal{F}_h\}_{h\in[H]}$, reference policy π^{ref}, $\alpha, \tau > 0$, mixture weight $\omega \in (0,1)$, OMD parameter $\rho \in (0,1)$, regularization parameter $\lambda > 0$, confidence radius $\beta > 0$, and number of episodes K.
2: Choose $r_1 \in \mathcal{R}$.
3: **for** $k = 1, \dots, K$ **do**
4: For each $h \in [H]$, define $u_{k,h} := u_{r_k,h}$, and set $\widehat{V}_{k,H+1} \equiv 0$.
5: **for** $h = H, H-1, \dots, 1$ **do**
6: $\widehat{f}_{k,h} \in \arg\min_{f\in\mathcal{F}_h} \sum_{j=1}^{k-1} \left(f(s_{j,h}, a_{j,h}) - u_{k,h}(s_{j,h}, a_{j,h}) - \widehat{V}_{k,h+1}(s_{j,h+1}) \right)^2$.
7: $b_{k,h}(s,a) = \min\left\{4V_{\max},\ \beta\, D_{\mathcal{F}_h}\left((s,a); \mathcal{D}_{k-1,h}; \lambda\right)\right\}$.
8: $\widehat{Q}_{k,h}(s,a) = \text{clip}_{[-V_{\max}, V_{\max}]}\left(\widehat{f}_{k,h}(s,a) + b_{k,h}(s,a)\right)$.
9: $\widehat{V}_{k,h}(s) = \tau \log \mathbb{E}_{a\sim\pi_h^{\text{ref}}(\cdot|s)}\left[\exp\left(\frac{\widehat{Q}_{k,h}(s,a)}{\tau}\right)\right]$.
10: $\pi_{k,h}(a \mid s) = \pi_h^{\text{ref}}(a \mid s) \exp\left(\frac{\widehat{Q}_{k,h}(s,a) - \widehat{V}_{k,h}(s)}{\tau}\right)$.
11: **end for**
12: Execute π_k for one episode and observe $\{(s_{k,h}, a_{k,h}, s_{k,h+1})\}_{h=1}^H$.
13: Form the observed reward gradient in Eq. 7 and compute r_{k+1} according to Eq. 6.
14: **end for**
15: **return** $\bar{\pi}_K = \text{Unif}\{\pi_1, \dots, \pi_K\}$, where one policy is sampled uniformly at the beginning of each evaluation episode.

Remark 1. *The generalized Eluder dimension in Definition 1 corresponds to the unweighted case $\sigma \equiv 1$ of Agarwal et al. (2023). It is controlled by the standard Eluder dimension $\dim_E(\mathcal{F}, \epsilon)$ (Russo & Roy, 2013), up to logarithmic factors and regularization-dependent terms (Zhao et al., 2024).*

In addition to sequential uncertainty, we quantify function-class size through covering numbers, which enter the uniform concentration bounds in our analysis.

Definition 2 (Covering numbers)**.** *For $\kappa > 0$ and a function class $\mathcal{C}$, let $\mathcal{N}_{\mathcal{C}}(\kappa)$ denote the smallest cardinality of a κ-cover $\mathcal{C}_\kappa \subseteq \mathcal{C}$: for every $f \in \mathcal{C}$, there exists $f' \in \mathcal{C}_\kappa$ such that $\|f - f'\|_\infty \le \kappa$.*

4 Dually Regularized Adversarial Imitation Learning

In this section, we introduce Dually Regularized AIL, summarized in Algorithm 1. The algorithm takes as input an expert dataset $\mathcal{D}^E$ consisting of N length-H trajectories generated by the expert policy π^E, together with a reference policy π^{ref}. We consider two general function classes: $\mathcal{F}$ for approximating the action-value functions and $\mathcal{R}$ for learning the adversarial reward.

Dually Regularized AIL proceeds for K episodes, each consisting of two phases. In the first phase, analogous to the standard KL-regularized RL (Zhao et al., 2025), the algorithm performs backward least-squares value estimation on a regularized reward, constructs exploration bonuses from the resulting confidence bounds, and updates the policy using the closed-form solution to the KL-regularized policy optimization problem (Lines 6–10). The resulting policy is then executed in the environment to collect a new trajectory (Line 12). In the second phase, the adversarial reward is updated using online mirror descent based on the expert and the behavioral episode from current policy (Line 13). We describe these two phases in detail below.

Phase I: Backward Planning and Least-Squares Regression. For fixed r_k, we incorporate the reward regularization in $\psi_{\pi_k}^{\omega}$ with the shaped reward by $u_{r,h}(s,a) := r_h(s,a) + \frac{1}{2}\cdot\alpha(1-\omega)r_h(s,a)^2$. Notably, by the definition of the regularized AIL objective Eq. 1, we have

$$\mathcal{L}(\pi, r) = J_{u_r}^{\tau}(\pi^E) - J_{u_r}^{\tau}(\pi) - \tfrac{1}{2}\cdot\alpha\textstyle\sum_{h=1}^H \mathbb{E}_{d_h^E}[r_h^2].$$

Thus, for fixed r, minimizing $\mathcal{L}(\pi, r)$ over π is equivalent to maximizing $J_{u_r}^{\tau}(\pi)$.

Then starting from $\widehat{V}_{k,H+1} \equiv 0$, we estimate the value function by fitting the Bellman targets over $\mathcal{F}_h$ using all previously collected transitions:

$$\widehat{f}_{k,h} \in \arg\min_{f \in \mathcal{F}_h} \sum_{j=1}^{k-1} \left(f(s_{j,h}, a_{j,h}) - u_{k,h}(s_{j,h}, a_{j,h}) - \widehat{V}_{k,h+1}(s_{j,h+1}) \right)^2 . \tag{3}$$

The targets are recomputed using the current shaped reward and continuation value, allowing historical transitions to support planning under changing rewards and policies without estimating the transition dynamics. We then add the exploration bonus

$$b_{k,h}(s,a) := \min\left\{4V_{\max},\ \beta D_{\mathcal{F}_h}((s,a); \mathcal{D}_{k-1,h}; \lambda)\right\},$$

where $\mathcal{D}_{k-1,h} := \{(s_{j,h}, a_{j,h})\}_{j=1}^{k-1}$ denotes the history of trajectories at step h before the episode k.

The algorithm clips the resulting action-value estimates. On the confidence event proved in Lemma 14, these estimates are optimistic and encourage exploration of uncertain state–action pairs. Finally, the soft Bellman update yields $\widehat{V}_{k,h}$ and the closed-form policy $\pi_{k,h}$, balancing optimistic action values against KL deviation from π^{ref}. We execute π_k for one episode and retain the trajectory for the reward update in Phase II and subsequent planning.

Phase II: Adversarial Reward Learning. For fixed π_k, the reward learner therefore seeks to control regret with respect to the following population loss

$$\ell_k(r) := \sum_{h=1}^{H} \left\{ \mathbb{E}_{d_h^{\pi_k}} \left[r_h + \tfrac{1}{2} \cdot \alpha(1-\omega) r_h^2 \right] - \mathbb{E}_{d_h^E} \left[r_h - \tfrac{1}{2} \cdot \alpha\omega r_h^2 \right] \right\}. \tag{4}$$

By the unknown nature of the environment and expert policy, the reward learner, however, is only exposed to the following empirical loss. After observing episode k, let $\hat{d}_{k,h} := \delta_{(s_{k,h}, a_{k,h})}$, $\hat{d}_h^E := \frac{1}{N} \sum_{i=1}^{N} \delta_{(s_h^i, a_h^i)}$, where $\delta_{(s,a)} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ denotes the one-hot vector at (s,a). The empirical loss is

$$\widehat{\ell}_k(r) := \sum_{h=1}^{H} \left\{ \mathbb{E}_{\hat{d}_{k,h}} \left[r_h + \tfrac{1}{2} \cdot \alpha(1-\omega) r_h^2 \right] - \mathbb{E}_{\hat{d}_h^E} \left[r_h - \tfrac{1}{2} \cdot \alpha\omega r_h^2 \right] \right\}. \tag{5}$$

The first expectation is evaluation at $(s_{k,h}, a_{k,h})$; the second is the average over the N expert trajectories. ℓ_k is used in the analysis, while $\widehat{\ell}_k$ determines the reward update.

After observing the loss feedback, we update the reward using the following online mirror descent (OMD) update rule with parameter $\rho \in (0,1)$:

$$r_{k+1} \in \arg\min_{r \in \mathcal{R}} \left\{ \langle \widehat{g}_k, r \rangle + \tfrac{1}{2} \cdot \alpha\rho \sum_{h=1}^{H} \|r_h - r_{k,h}\|_{k,h}^2 \right\}. \tag{6}$$

where $\widehat{g}_k := (\widehat{g}_{k,h})_{h=1}^{H}$ is the observed loss gradient vector at r_k, with components

$$\widehat{g}_{k,h} := (1 + \alpha(1-\omega) r_{k,h}) \circ \hat{d}_{k,h} - (1 - \alpha\omega r_{k,h}) \circ \hat{d}_h^E , \tag{7}$$

for $\circ$ denoting the elementwise multiplication, and $\|r_h - r_{k,h}\|_{k,h}^2$ is the Bregman divergence between r_h and $r_{k,h}$ with respect to the following squared seminorm function

$$\|r\|_{k,h}^2 := \frac{k\omega}{N} \sum_{i=1}^{N} r(s_h^i, a_h^i)^2 + (1-\omega) \sum_{j=1}^{k} r(s_{j,h}, a_{j,h})^2, \quad \forall r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}. \tag{8}$$

We note that no-regret learning algorithms such as follow-the-regularized-leader (FTRL) and online mirror descent (OMD) (Hazan, 2023) have also been used for policy learning (Sun et al., 2019), reward learning (Li et al., 2026; Xu et al., 2024), or both (Shani et al., 2022) in many other AIL works. Our approach specializes this approach with an adaptive proximal penalty choice, which aligns with our choice of reward penalty with only an extra multiplier ρ. This design allows us to derive a reward class Eluder dimension bound for a stability term derived from the OMD analysis and offset part of the sampling error. These two benefits together help us to provide a sharp bound of the reward error in later analysis. This choice also falls within the general framework of adaptive mirror descent discussed in McMahan (2017).

5 Theoretical Analysis

In this section, we present the sharp regret and sample complexity guarantee for Algorithm 1. We start by several common assumptions related to the general function approximation:

Definition 3 (Bonus classes). *For each $h \in [H]$, let $\mathcal{B}_h$ be a nonempty deterministic class of functions from $\mathcal{S} \times \mathcal{A}$ to $[0, 4V_{\max}]$ containing every bonus $b_{k,h}(s,a) = \min\{4V_{\max}, \beta D_{\mathcal{F}_h}((s,a); \mathcal{D}_{k-1,h}; \lambda)\}$ generated by Algorithm 1, over all episodes and possible data histories at the prescribed deterministic planner parameters.*

We further assume the reward class $\mathcal{R}$, value function class $\mathcal{F}$ and its bonus class $\mathcal{B}$ have finite covering numbers $\overline{\mathcal{N}}_{\mathcal{R}}(\kappa), \overline{\mathcal{N}}_{\mathcal{F}}(\kappa), \overline{\mathcal{N}}_{\mathcal{B}}(\kappa)$ under any $\kappa > 0$. then we assume the Bellman completeness of the value estimation for any reward under general function approximation which is a common assumption for the theoretical analysis in model-free RL (Agarwal et al., 2023; Zhao et al., 2024).

Assumption 1 (Bellman completeness). *For every $h \in [H]$, $r \in \mathcal{R}$, and $V \in \mathcal{V}_{h+1}$ from the optimistic soft-value class in Definition 4, $u_{r,h} + P_h V \in \mathcal{F}_h$. The terminal continuation is $V \equiv 0$.*

With these assumption for completeness, we are ready to present our theorem.

Theorem 1. *Under the stated assumptions, for any $\delta \in (0,1)$ and $\omega \in (0,1)$, set parameter as $\rho = 1/2$, $\lambda > 0$, and $\beta := \sqrt{\lambda + 64V_{\max}^2 \log \frac{8HK\overline{\mathcal{N}}_{\mathcal{F}\oplus\mathcal{B}}(V_{\max}/K^2)}{\delta}}$, with probability at least $1-\delta$,*

$$\begin{aligned}\mathrm{Regret}_{\omega,\alpha,\tau}(K) \le \mathcal{O}\Bigg(& \frac{H(1+\alpha\omega)^2}{\alpha\omega}\Big(1 + \log K + \frac{K}{N}\log\frac{4H\overline{\mathcal{N}}_{\mathcal{R}}(1/N^2)^2}{\delta}\Big) \\ & + \frac{H(1+\alpha(1-\omega))^2}{\alpha(1-\omega)}\Big(\dim_K(\mathcal{R}, 4/K^2) + \log\frac{4H\overline{\mathcal{N}}_{\mathcal{R}}(1/K^2)}{\delta}\Big) \\ & + \frac{H^3\beta^2}{\tau}\left[\dim_K(\mathcal{F},\lambda) + \log\frac{8H}{\delta}\right]\Bigg). \end{aligned} \tag{9}$$

where $\overline{\mathcal{N}}_{\mathcal{F}\oplus\mathcal{B}}(\kappa) := \max_{h\in[H]}\{\mathcal{N}_{\mathcal{F}_h}(\kappa)\mathcal{N}_{\mathcal{F}_{h+1}}(\frac{\kappa}{2})\mathcal{N}_{\mathcal{B}_{h+1}}(\frac{\kappa}{2})\}$, $\overline{\mathcal{N}}_{\mathcal{R}}(\kappa) := \max_{h\in[H]}\mathcal{N}_{\mathcal{R}_h}(\kappa)$

Theorem 1 immediately yields the following sample-complexity guarantee.

Corollary 1 (Sample complexity of Dually Regularized AIL). *Under the conditions of Theorem 1, suppose $\lambda \le V_{\max}^2$, and write $d_{\mathcal{R}} := \dim_K(\mathcal{R}, 4/K^2)$ and $d_{\mathcal{F}} := \max\{1, \dim_K(\mathcal{F},\lambda)\}$. For any $\epsilon > 0$, the returned policy mixture $\bar{\pi}_K = \mathrm{Unif}\{\pi_1,\dots,\pi_K\}$ and averaged reward $\bar{r}_K = K^{-1}\sum_{k=1}^K r_k$ satisfy, with probability at least $1-\delta$, the averaged dual gap satisfied*

$$0 \le \sup_{r\in\mathcal{R}} \mathcal{L}(\bar{\pi}_K, r) - \inf_{\pi} \mathcal{L}(\pi, \bar{r}_K) \le \epsilon$$

with sufficient interaction and expert sample requirements

$$\begin{aligned} K &= \widetilde{\mathcal{O}}\Bigg(\frac{H}{\epsilon}\left[\frac{(1+\alpha\omega)^2}{\alpha\omega} + \frac{(1+\alpha(1-\omega))^2}{\alpha(1-\omega)}\left(d_{\mathcal{R}} + \log\big(2\overline{\mathcal{N}}_{\mathcal{R}}(1/K^2)\big)\right)\right] \\ &\qquad + \frac{H^3V_{\max}^2 d_{\mathcal{F}}}{\tau\epsilon}\log\big(2\overline{\mathcal{N}}_{\mathcal{F}\oplus\mathcal{B}}(V_{\max}/K^2)\big)\Bigg), \\ N &= \widetilde{\mathcal{O}}\left(\frac{H(1+\alpha\omega)^2}{\alpha\omega\epsilon}\log\big(2\overline{\mathcal{N}}_{\mathcal{R}}(1/N^2)\big)\right). \end{aligned}$$

These are implicit sufficient conditions because the complexities and covering scales can depend on K and N. Sample counts are rounded up to positive integers. The notation $\widetilde{\mathcal{O}}$ suppresses additional logarithmic factors in H, K, and $1/\delta$.

Remark 2 (Interpretation of the sample complexity). *For fixed $\alpha, \tau > 0$ and $\omega \in (0,1)$, we have $V_{\max} = \Theta(H)$. Corollary 1 therefore yields a policy-planning contribution of $\widetilde{\mathcal{O}}\left(\tau^{-1}H^5 d_{\mathcal{F}}\log\big(\overline{\mathcal{N}}_{\mathcal{F}\oplus\mathcal{B}}(V_{\max}/K^2)\big)/\epsilon\right)$ to the required number of online interactions, while the expert-trajectory requirement is $\mathcal{O}\left(H\log\big(\overline{\mathcal{N}}_{\mathcal{R}}(1/N^2)/\delta\big)/\epsilon\right)$. For comparison, after translating the KL-LSVI-UCB result of Zhao et al. (2025) to our notation and horizon convention, its interaction complexity is $\widetilde{\mathcal{O}}\left(\tau^{-1}H^5 d_{\mathcal{F}}\log\big(\overline{\mathcal{N}}_{\mathcal{F}\oplus\mathcal{B}}(V_{\max}/K^2)\big)/\epsilon\right)$. Thus, the policy-planning component of Algorithm 1 matches the interaction-complexity order of KL-LSVI-UCB.*

Remark 3. *Regarding the dependence on the precision ϵ, our algorithm achieves an expert sample complexity of $\widetilde{\mathcal{O}}(\epsilon^{-1})$ for both deterministic and stochastic experts, with fixed regularization parameters and controlled function-class complexity. In contrast, for the unregularized imitation gap under general policy classes, Foster et al. (2024) establish worst-case upper bounds of $\widetilde{\mathcal{O}}(\epsilon^{-1})$ for deterministic experts and $\widetilde{\mathcal{O}}(\epsilon^{-2})$ for stochastic experts, together with matching $\Omega(\epsilon^{-1})$ and $\Omega(\epsilon^{-2})$ lower bounds in their dependence on ϵ. Similarly, Li et al. (2026) derive an interaction lower bound for unregularized AIL containing an $\epsilon^{-2}\exp(-N)$ term, where N is the number of expert trajectories. Our faster expert and interaction rates do not contradict these results: our guarantee controls the regularized saddle-point gap for fixed positive reward and policy regularization, rather than the unregularized imitation gap considered in those works. Consequently, these rates do not directly imply the same sample complexity for the unregularized objective, for which regularization bias and the dependence on the regularization parameters must also be accounted for.*

Remark 4 (Role of reward regularization)**.** *The quadratic regularizer ψ_π^ω controls both reward-update stability and sampling error. Matching the proximal penalty to the empirical curvature yields an OMD stability bound governed by the generalized eluder dimension of $\mathcal{R}$, while $\rho \in (0,1)$ reserves curvature to absorb both expert and learner sampling errors.*

The weight ω allocates curvature between expert and learner occupancies; approaching either endpoint weakens the corresponding error bound. The symmetric choice $\omega = 1/2$ matches the equal-mixture regularizer in LS-IQ (Al-Hafez et al., 2023), which admits a bounded Pearson χ^2-divergence interpretation. Our analysis complements this perspective with finite-sample guarantees: together with KL-regularized planning, any fixed $\omega \in (0,1)$ preserves the $\widetilde{\mathcal{O}}(1/K + 1/N)$ regularized gap bound for fixed regularization parameters and controlled function-class complexity.

6 Proof Sketch

We sketch the proof of Theorem 1, emphasizing how reward curvature controls both optimization and estimation, while policy curvature controls planning uncertainty. We suppress dependence on the fixed reward parameters α, ω, ρ and additional logarithmic factors; Eq. 9 and Appendix C give the full bounds and proofs.

By Lemma 6, the cumulative regret separates into reward-learning and policy-planning errors:

$$\mathrm{Regret}_{\omega,\alpha,\tau}(K) \le \underbrace{\sup_{r\in\mathcal{R}} \sum_{k=1}^K \left[\ell_k(r_k) - \ell_k(r)\right]}_{T_1:\text{ reward-learning error}} + \underbrace{\sum_{k=1}^K \left[V^{\star,\tau}_{u_k,1}(s_1) - J^\tau_{u_k}(\pi_k)\right]}_{T_2:\text{ policy-planning error}}.$$

Then T_1 could be further decomposed into the empirical loss on $\hat{\ell}_k$ and the concentration between the expected loss ℓ_k and empirical loss $\hat{\ell}_k$:

$$T_1 = \sup_{r\in\mathcal{R}} \left(\sum_{k=1}^K \left[\widehat{\ell}_k(r_k) - \widehat{\ell}_k(r)\right] + \sum_{k=1}^K \left[(\ell_k - \widehat{\ell}_k)(r_k) - (\ell_k - \widehat{\ell}_k)(r)\right] \right)$$

The difficulty in controlling T_1 is that the reward iterates depend on both sequential learner feedback and the repeatedly reused expert dataset. Rather than treating the empirical gradient as conditionally unbiased, we retain curvature to control the second term inside the supremum above, which is detailed in the following lemma:

Lemma 1 (Eluder-controlled stability with retained curvature, informal)**.** *The curvature-matched OMD update ensures, for every $r \in \mathcal{R}$,*

$$\sum_{k=1}^K \left[\widehat{\ell}_k(r_k) - \widehat{\ell}_k(r)\right] \le \mathcal{O}\big(H\left[1 + \log K + \dim_K(\mathcal{R}, 4/K^2)\right]\big) - \tfrac{1}{2}\cdot\alpha(1-\rho)\mathcal{E}_K(r),$$

where $\mathcal{E}_K(r)$ is the loss curvature enduced by regularization, i.e.,

$$\mathcal{E}_K(r) := \sum_{k=1}^K \sum_{h=1}^H \left[\omega \mathbb{E}_{\hat{d}_h^E}[(r_{k,h} - r_h)^2] + (1-\omega)\mathbb{E}_{\hat{d}_{k,h}}[(r_{k,h} - r_h)^2]\right].$$

This bound holds over the general convex reward class, without requiring the empirical reward seminorm to be nondegenerate.

The proximal penalty is ρ times the cumulative empirical loss curvature. The OMD three-point inequality therefore telescopes in this data-dependent seminorm while preserving the $(1-\rho)$ fraction

of comparator curvature. Repeated expert observations contribute a harmonic sum, whereas learner updates are controlled by ratios of current squared reward changes to their accumulated magnitude. These ratios yield the generalized eluder dimension of $\mathcal{R}$ as detailed in Lemmas 9 and 10.

Lemma 2 (Variance cancellation under adaptive expert-data reuse, informal)**.** *With probability at least* $1-\delta/2$*, simultaneously for all* $r \in \mathcal{R}$*,*

$$\sum_{k=1}^{K}\big[(\ell_k-\widehat{\ell}_k)(r_k)-(\ell_k-\widehat{\ell}_k)(r)\big] \le \tfrac{1}{2}\cdot\alpha(1-\rho)\mathcal{E}_K(r) + \widetilde{\mathcal{O}}\Big((K+N)H/N\cdot\big(\log\big(2\overline{\mathcal{N}}_{\mathcal{R}}(1/K^2)\big)+\log\big(2\overline{\mathcal{N}}_{\mathcal{R}}(1/N^2)\big)\big)\Big).$$

The expert-data bound is uniform over reward pairs, allowing every iterate r_k *to depend on the same expert sample.*

Our concentration argument adds a centered quadratic correction to the sampling noise before applying Freedman's inequality. Its population variance is then canceled by the corresponding population quadratic term, leaving precisely the empirical curvature appearing above. For learner feedback, we apply this argument sequentially and extend uniformly from a $1/K^2$-cover of the comparator class. For expert feedback, we first establish concentration over all pairs in a $1/N^2$-cover, and only then substitute the data-dependent iterate r_k for one member of the pair. This order avoids an independence assumption between the learned reward and the demonstrations; see Lemma 12.

Adding the two inequalities cancels $\mathcal{E}_K(r)$ exactly, before taking the supremum over r. Choosing $\rho = 1/2$ balances the $1/\rho$ stability and $1/(1-\rho)$ concentration factors, reserving a constant fraction of curvature for each argument. This gives Lemma 7: the online reward cost is controlled by sequential complexity and covering logarithms, while expert reuse contributes $\widetilde{\mathcal{O}}(KH/N)$. Thus, averaging yields a $1/N$ expert-data contribution without requiring a deterministic expert or fresh demonstrations at each episode.

For T_2, we build on the KL-regularized planning analysis of Zhao et al. (2025), accounting for the changing shaped rewards and data-dependent continuation values.

Lemma 3 (Squared-bonus under changing rewards, informal)**.** *With probability at least* $1-\delta/2$*,*

$$T_2 \le \frac{2H^2}{\tau}\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{d_h^{\pi_k}}\big[b_{k,h}^2\big] \le \mathcal{O}\left(\frac{H^3\beta^2}{\tau}\left[\dim_K(\mathcal{F},\lambda)+\log\frac{8H}{\delta}\right]\right).$$

Uniform regression confidence ensures optimism despite reuse of historical transitions with recomputed targets. The KL-regularized performance difference identity and Gibbs formulas then make policy error quadratic in the optimistic action-value error. A Bellman recursion reduces this error to squared bonuses. Finally, concentration transfers the observed squared-bonus bound, controlled by $\dim_K(\mathcal{F},\lambda)$, to occupancy expectations. This quadratic dependence, rather than a linear bonus bound, gives the fast planning rate; see Lemma 8.

Combining the reward and policy bounds of T_1, T_2 presented in aforementioned lemmas, we are ready to control the cumulative dual regret in Eq. 9 by

$$\begin{aligned}\mathrm{Regret}_{\omega,\alpha,\tau}(K) \le{}& \tilde{\mathcal{O}}(H(1+\log K+\dim_K(\mathcal{R},4/K^2))) - \tfrac{1}{2}\cdot\alpha(1-\rho)\mathcal{E}_K(r) && \text{(Lemma 1)}\\ &+\tfrac{1}{2}\cdot\alpha(1-\rho)\mathcal{E}_K(r)+\tilde{\mathcal{O}}((K+N)\cdot H/N\cdot\log\overline{\mathcal{N}}_{\mathcal{R}}) && \text{(Lemma 2)}\\ &+\tilde{\mathcal{O}}(H^3\beta^2\dim_K(\mathcal{F},\lambda)/\tau), && \text{(Lemma 3)}\end{aligned}$$

which yields the results in Theorem 1. Averaging over episodes gives the $\widetilde{\mathcal{O}}(1/K+1/N)$ regularized saddle-point gap for fixed regularization parameters and controlled function-class complexity.

7 Conclusion

In this paper, we considered the provable benefits of reward and policy regularizers in adversarial imitation learning setting. We proposed Dually Regularized AIL algorithm that minimizes a regularized imitation learning gap. Our algorithm establishes an $\widetilde{\mathcal{O}}(\frac{1}{K}+\frac{1}{N})$ average regret bound under fixed regularization. This result demonstrates the benefits of regularization where the reward regularization absorbs sampling error while controlling update stability, and policy regularization controls the planner error, which together contributes to a faster convergence rate for AIL.

AI USE STATEMENT

We used generative AI tools to assist with refining mathematical proofs and revising the manuscript for clarity and presentation. All AI-assisted content, including proof derivations and textual revisions, was carefully reviewed and verified by the authors for correctness and consistency with the underlying results. The authors take full responsibility for the final content of this work.

REFERENCES

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, pp. 1, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138385. doi: 10.1145/1015330.1015430.

Alekh Agarwal, Yujia Jin, and Tong Zhang. VOQL: Towards optimal regret in model-free RL with nonlinear function approximation. In *Proceedings of the Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 987–1063. PMLR, 2023.

Firas Al-Hafez, Davide Tateo, Oleg Arenz, Guoping Zhao, and Jan Peters. Ls-iq: Implicit reward regularization for inverse reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Dylan J Foster, Adam Block, and Dipendra Misra. Is behavior cloning all you need? understanding horizon in imitation learning. *Advances in Neural Information Processing Systems*, 37:120602–120666, 2024.

Divyansh Garg, Shuvam Chakraborty, Chris Cundy, Jiaming Song, and Stefano Ermon. IQ-Learn: Inverse soft-Q learning for imitation. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/210f760a89db30aa72ca258a3483cc7f-Abstract.html.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Elad Hazan. *Introduction to Online Convex Optimization*. arXiv, August 2023. URL http://arxiv.org/abs/1909.05207. arXiv:1909.05207 [cs].

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/cc7e2b878868cbae992d1fb743995d8f-Abstract.html.

Rohit Jena, Changliu Liu, and Katia Sycara. Augmenting gail with bc for sample efficient imitation learning. In *Conference on Robot Learning*, pp. 80–90. PMLR, 2021.

Shangzhe Li, Dongruo Zhou, and Weitong Zhang. Near-optimal second-order guarantees for model-based adversarial imitation learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/dda456ad57133d709520702851cc320e-Paper-Conference.pdf.

Zhihan Liu, Yufeng Zhang, Zuyue Fu, Zhuoran Yang, and Zhaoran Wang. Provably efficient generative adversarial imitation learning for online and offline setting with linear function approximation. *arXiv preprint arXiv:2108.08765*, 2021.

H. Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *J. Mach. Learn. Res.*, 18(1):3117–3166, January 2017. ISSN 1532-4435.

Anupam Nayak, Tong Yang, Osman Yağan, Gauri Joshi, and Yuejie Chi. Achieving logarithmic regret in KL-regularized zero-sum markov games, 2025. URL https://arxiv.org/abs/2510.13060. Revised arXiv version, 2026.

Manu Orsini, Anton Raichuk, Léonard Hussenot, Damien Vincent, Robert Dadashi, Sertan Girgin, Matthieu Geist, Olivier Bachem, Olivier Pietquin, and Marcin Andrychowicz. What matters for adversarial imitation learning? *Advances in Neural Information Processing Systems*, 34:14656–14668, 2021.

Nived Rajaraman, Lin Yang, Jiantao Jiao, and Kannan Ramchandran. Toward the fundamental limits of imitation learning. *Advances in Neural Information Processing Systems*, 33:2914–2924, 2020.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2256–2264. Curran Associates, Inc., 2013.

Lior Shani, Tom Zahavy, and Shie Mannor. Online apprenticeship learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 8240–8248, 2022.

Wen Sun, Anirudh Vemula, Byron Boots, and Drew Bagnell. Provably efficient imitation learning from observation alone. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 6036–6045. PMLR, May 2019. URL https://proceedings.mlr.press/v97/sun19b.html.

Umar Syed and Robert E Schapire. A game-theoretic approach to apprenticeship learning. In J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

Daniil Tiapkin, Denis Belomestny, Daniele Calandriello, Eric Moulines, Remi Munos, Alexey Naumov, Pierre Perrault, Yunhao Tang, Michal Valko, and Pierre Menard. Fast rates for maximum entropy exploration. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 34161–34221. PMLR, 2023. URL https://proceedings.mlr.press/v202/tiapkin23a.html.

Daniil Tiapkin, Denis Belomestny, Daniele Calandriello, Eric Moulines, Aleksei Naumov, Pierre Perrault, Michal Valko, and Pierre Ménard. Demonstration-regularized RL. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/7b5d3047939b63ed97fcbbee23f8eb77-Abstract-Conference.html.

Luca Viano, Stratis Skoulakis, and Volkan Cevher. Imitation learning in discounted linear mdps without exploration assumptions. In *Forty-first International Conference on Machine Learning*, 2024.

Tian Xu, Ziniu Li, Yang Yu, and Zhi-Quan Luo. Provably efficient adversarial imitation learning with unknown transitions. In *Uncertainty in Artificial Intelligence*, pp. 2367–2378. PMLR, 2023.

Tian Xu, Zhilong Zhang, Ruishuo Chen, Yihao Sun, and Yang Yu. Provably and practically efficient adversarial imitation learning with general function approximation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 66108–66146, December 2024. URL https://proceedings.neurips.cc//paper_files/paper/2024/hash/79b90b4c2ee23cc35fdd8de2969dc4e8-Abstract-Conference.html.

Tong Zhang. *Mathematical Analysis of Machine Learning Algorithms*. Cambridge University Press, 2023. ISBN 978-1009098380.

Heyang Zhao, Jiafan He, and Quanquan Gu. A nearly optimal and low-switching algorithm for reinforcement learning with general function approximation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 94684–94735. Curran Associates, Inc., 2024. doi: 10.52202/079017-3002.

Heyang Zhao, Chenlu Ye, Wei Xiong, Quanquan Gu, and Tong Zhang. Logarithmic regret for online KL-regularized reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 77864–77884. PMLR, 2025.

A Additional Related Works

Fast rates in regularized reinforcement learning. Maximum-entropy RL provides a practical policy-regularization framework (Haarnoja et al., 2018). On the theoretical side, Tiapkin et al. (2023) establish fast rates for maximum-entropy exploration and entropy-regularized planning, and Tiapkin et al. (2024) study KL-regularized planning toward a reference policy learned from demonstrations. These best-policy-identification guarantees concern the returned policy, rather than cumulative regret during exploration. Zhao et al. (2025) establish logarithmic regret under general function approximation by combining optimistic value estimation with a squared-Bellman-error decomposition. Our policy analysis builds on this mechanism. Related logarithmic-regret results extend KL regularization to zero-sum Markov games (Nayak et al., 2025). Neither the fixed-reward RL problem nor the Markov-game protocol directly accounts for an adversarial reward learner that repeatedly reuses the same finite expert sample. Controlling this reward-learning error is the additional requirement addressed by our analysis.

Sequential complexity and reward optimization. Eluder dimension quantifies the sequential uncertainty of function classes (Russo & Roy, 2013); generalized versions support optimistic RL with nonlinear function approximation (Agarwal et al., 2023; Zhao et al., 2024; 2025). We use the unweighted scalar-ridge specialization of generalized eluder dimension for both reward-update stability and value estimation. The generalized eluder coefficient in OPT-AIL is a related but distinct complexity measure (Xu et al., 2024). Online mirror descent itself is standard (Hazan, 2023; Shani et al., 2022); our analysis uses the occupancy-weighted reward curvature to control sequential learner noise and finite-expert error while bounding the stability cost directly through the reward class, without assuming a well-conditioned expert feature covariance.

B Properties of KL-Regularized RL

In this section, we present some of the basic properties of KL-regularized RL.

Let $V^{\star,\tau}_{r,h} := \sup_\pi V^{\pi,\tau}_{r,h}$ and $Q^{\star,\tau}_{r,h} := \sup_\pi Q^{\pi,\tau}_{r,h}$ denote the pointwise optimal values. With $V^{\star,\tau}_{r,H+1} \equiv 0$, they satisfy

$$Q^{\star,\tau}_{r,h}(s,a) = r_h(s,a) + [P_h V^{\star,\tau}_{r,h+1}](s,a), \tag{10}$$

$$V^{\star,\tau}_{r,h}(s) = \tau \log \mathbb{E}_{a\sim\pi^{\mathrm{ref}}_h(\cdot|s)} \exp\left(\frac{Q^{\star,\tau}_{r,h}(s,a)}{\tau}\right). \tag{11}$$

Existing results on KL-regularized reinforcement learning (Zhang, 2023) yield a closed-form characterization of the optimal policy, as stated in the following lemma:

Lemma 4 (Optimal KL-regularized policy, Zhang (2023))**.** *For any bounded stagewise reward $r = \{r_h\}_{h\in[H]}$, the policy $\pi^{\star,\tau}_r = \{\pi^{\star,\tau}_{r,h}\}_{h\in[H]}$ defined by*

$$\pi^{\star,\tau}_{r,h}(a \mid s) := \pi^{\mathrm{ref}}_h(a \mid s) \exp\left(\frac{Q^{\star,\tau}_{r,h}(s,a) - V^{\star,\tau}_{r,h}(s)}{\tau}\right) \tag{12}$$

attains the optimal value $V^{\star,\tau}_{r,h}(s)$ for every $h \in [H]$ and $s \in \mathcal{S}$.

The following identity expresses a policy's regularized value gap as its expected cumulative KL divergence from the optimal policy.

Lemma 5 (Performance difference for KL-regularized RL)**.** *For any bounded stagewise reward r and any policy π with finite expected cumulative KL cost relative to π^{ref}, we have*

$$V^{\star,\tau}_{r,1}(s_1) - J^\tau_r(\pi) = \tau \mathbb{E}_{P,\pi}\left[\sum_{h=1}^{H} \mathrm{KL}\Big(\pi_h(\cdot \mid s_h) \,\|\, \pi^{\star,\tau}_{r,h}(\cdot \mid s_h)\Big)\right], \tag{13}$$

where $\pi^{\star,\tau}_r$ is defined in Lemma 4.

Proof. On state–action pairs visited by π, Eq. 12 gives, almost surely,

$$\tau \log \frac{\pi_h(a \mid s)}{\pi^{\star,\tau}_{r,h}(a \mid s)} = \tau \log \frac{\pi_h(a \mid s)}{\pi^{\mathrm{ref}}_h(a \mid s)} - Q^{\star,\tau}_{r,h}(s,a) + V^{\star,\tau}_{r,h}(s).$$

Taking expectations along a trajectory under π and using the Bellman equations gives, for each $h \in [H]$,

$$\begin{aligned}
&\mathbb{E}_{P,\pi}\left[V^{\star,\tau}_{r,h}(s_h) - V^{\pi,\tau}_{r,h}(s_h)\right] \\
&\quad = \tau \mathbb{E}_{P,\pi}\left[\mathrm{KL}\Big(\pi_h(\cdot \mid s_h) \,\|\, \pi^{\star,\tau}_{r,h}(\cdot \mid s_h)\Big)\right] \\
&\qquad + \mathbb{E}_{P,\pi}\left[V^{\star,\tau}_{r,h+1}(s_{h+1}) - V^{\pi,\tau}_{r,h+1}(s_{h+1})\right].
\end{aligned}$$

Summing over h telescopes the value differences. The result follows from $V^{\star,\tau}_{r,H+1} = V^{\pi,\tau}_{r,H+1} = 0$ and $J^\tau_r(\pi) = V^{\pi,\tau}_{r,1}(s_1)$. □

C Proof of Theorem 1

Proof conventions. The expert has finite expected KL cost by assumption, and the algorithm's Gibbs policies also have finite expected KL cost, so the payoff differences below are well defined.

All $\mathcal{O}(\cdot)$ notation hides only universal numerical constants. Let $\mathcal{H}_{k-1}$ be the σ-algebra generated by the history before episode k, including $\mathcal{D}^E$ and the choices of r_k and π_k. Let $\mathcal{H}_{k,h}$ additionally include the observations in episode k up to $(s_{k,h}, a_{k,h})$, before $s_{k,h+1}$ is observed. Conditional on $\mathcal{H}_{k-1}$, $(s_{k,h}, a_{k,h})$ has distribution $d^{\pi_k}_h$; conditional on $\mathcal{H}_{k,h}$, $s_{k,h+1}$ has distribution $P_h(\cdot \mid s_{k,h}, a_{k,h})$. No independence between stages is assumed. Within the learner-sampling, regression, and exploration-bonus proofs, probabilities are conditional on $\mathcal{D}^E$; their bounds also hold unconditionally by averaging over $\mathcal{D}^E$.

C.1 Main Proof of Theorem 1

In this subsection, we highlight the key ingredients in the proof of Theorem 1. We use T_1 and T_2 from Eq. 14 for the reward-learning and policy-planning errors, respectively. Throughout this proof, the assumptions and parameter choices of Theorem 1 are in force.

Lemma 6 (Regret decomposition). *For every generated sequence, with $\bar\pi_K = \mathrm{Unif}\{\pi_1, \dots, \pi_K\}$ and $\bar r_K = K^{-1}\sum_{k=1}^K r_k$, we have*

$$\mathrm{Regret}_{\omega,\alpha,\tau}(K) \le \underbrace{\sup_{r \in \mathcal{R}} \sum_{k=1}^K \big(\ell_k(r_k) - \ell_k(r)\big)}_{T_1\text{: reward-learning error}} + \underbrace{\sum_{k=1}^K \left[V^{\star,\tau}_{u_k,1}(s_1) - J^\tau_{u_k}(\pi_k)\right]}_{T_2\text{: policy-planning error}}. \tag{14}$$

Here ℓ_k is the loss already defined in Eq. 4.

Proof of Lemma 6. Adding and subtracting $\sum_{k=1}^K \mathcal{L}(\pi_k, r_k)$ gives

$$\mathrm{Regret}_{\omega,\alpha,\tau}(K) = T_1 + \sum_{k=1}^K \mathcal{L}(\pi_k, r_k) - \inf_\pi \sum_{k=1}^K \mathcal{L}(\pi, r_k), \tag{15}$$

where Eq. 4 gives $\mathcal{L}(\pi_k, r) - \mathcal{L}(\pi_k, r_k) = \ell_k(r_k) - \ell_k(r)$.

By the definition of shaped-reward definition $u_{r,h}$ in Algorithm 1 $\mathcal{L}(\pi, r_k) = J^\tau_{u_k}(\pi^E) - J^\tau_{u_k}(\pi) - \frac{\alpha}{2}\sum_{h=1}^H \mathbb{E}_{d^E_h}[r^2_{k,h}]$. The terms involving the expert are independent of π, so

$$\begin{aligned}
\sum_{k=1}^K \mathcal{L}(\pi_k, r_k) - \inf_\pi \sum_{k=1}^K \mathcal{L}(\pi, r_k) &\le \sum_{k=1}^K \Big[\mathcal{L}(\pi_k, r_k) - \inf_\pi \mathcal{L}(\pi, r_k)\Big] \\
&= \sum_{k=1}^K \left[V^{\star,\tau}_{u_k,1}(s_1) - J^\tau_{u_k}(\pi_k)\right] = T_2.
\end{aligned} \tag{16}$$

Combining this with Eq. 15 proves Eq. 14. □

The first term T_1 in Eq. 14 controls reward learning error. We control it through the following lemma.

Lemma 7 (Reward-learning error). *Under the parameter choice $\rho = 1/2$ in Eq. 6, with probability at least $1 - \delta/2$,*

$$T_1 \le \mathcal{O}\Bigg(\frac{H(1+\alpha\omega)^2}{\alpha\omega}(1+\log K) + \frac{H(1+\alpha(1-\omega))^2}{\alpha(1-\omega)} \dim_K(\mathcal{R}, 4/K^2) + \frac{H(1+\alpha(1-\omega))^2}{\alpha(1-\omega)} \log \frac{4H\overline{\mathcal{N}}_{\mathcal{R}}(1/K^2)}{\delta} + \frac{KH(1+\alpha\omega)^2}{\alpha\omega N} \log \frac{4H\overline{\mathcal{N}}_{\mathcal{R}}(1/N^2)^2}{\delta} \Bigg).$$

Proof Sketch of Lemma 7. Combining the three-point inequality for the reward update with the exact quadratic expansion of $\widehat{\ell}_k$, Lemma 9 gives

$$T_1 \le T_S + \sup_{r \in \mathcal{R}} T_C(r).$$

Here T_S is the stability term in Eq. 17, and $T_C(r)$ is the sampling term with retained comparator curvature in Eq. 18.

The quadratic loss curvature offsets growth of the proximal penalty, leaving squared-update control in T_S and comparator curvature in $T_C(r)$. Stability is controlled by accumulated reward discrepancies on observed state–action pairs; Lemma 10 gives

$$T_S \le \frac{H(1+\alpha\omega)^2}{2\alpha\rho\omega}(1+\log K) + \frac{H(1+\alpha(1-\omega))^2}{2\alpha\rho(1-\omega)} \dim_K(\mathcal{R}, 4/K^2) + 2H(1+\alpha(1-\omega)).$$

For sampling error, the retained comparator curvature offsets the variance terms in Freedman's inequality. Uniform covering arguments then give, by Lemma 12, with probability at least $1 - \delta/2$,

$$\sup_{r \in \mathcal{R}} T_C(r) \le \mathcal{O}\Bigg(\frac{H(1+\alpha(1-\omega))^2}{\alpha(1-\omega)(1-\rho)} \log \frac{4H\overline{\mathcal{N}}_{\mathcal{R}}(1/K^2)}{\delta} + \frac{KH(1+\alpha\omega)^2}{\alpha\omega(1-\rho)N} \log \frac{4H\overline{\mathcal{N}}_{\mathcal{R}}(1/N^2)^2}{\delta} \Bigg).$$

Combining the stability and concentration bounds and setting $\rho = 1/2$ proves the claim. The detailed argument is given in Appendix C.2. □

Remark 5 (Choice of the OMD parameter). *We choose $\rho = 1/2$ for concreteness. The analysis extends to any $\rho \in (0,1)$, with the stability and sampling-error bounds carrying factors $1/\rho$ and $1/(1-\rho)$, respectively. Hence any fixed $\rho \in (0,1)$ yields the same dependence on the sample budgets, with constants depending on ρ.*

We next control the policy-planning error T_2 with the following lemma.

Lemma 8 (Policy-planning error). *With probability at least $1 - \delta/2$,*

$$T_2 \le \frac{2H^2}{\tau} \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi_k}}[b_{k,h}^2] \le \mathcal{O}\left(\frac{H^3\beta^2}{\tau} \left[\dim_K(\mathcal{F}, \lambda) + \log \frac{8H}{\delta} \right] \right).$$

Proof Sketch of Lemma 8. The analysis starts from the optimism property $\widehat{Q}_{k,h}(s,a) \ge Q^{\star,\tau}_{u_k,h}(s,a)$ in Lemma 15, which holds simultaneously for all $(k,h,s,a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$ with probability at least $1 - \delta/4$.

Under optimism, the formulas for the algorithm's policy and the optimal KL-regularized policy give

$$\tau\,\mathrm{KL}\Big(\pi_{k,h}(\cdot\mid s)\,\|\,\pi^{\star,\tau}_{u_k,h}(\cdot\mid s)\Big) \le \frac{1}{2\tau}\mathbb{E}_{a\sim\pi_{k,h}(\cdot|s)}\Big[\big(\widehat{Q}_{k,h}(s,a)-Q^{\star,\tau}_{u_k,h}(s,a)\big)^2\Big].$$

The KL-regularized performance difference identity therefore bounds the policy-planning error quadratically in the action-value estimation error. Controlling this estimation error through the exploration bonuses, Lemma 16 yields the squared-bonus bound in the statement.

The generalized eluder dimension of $\mathcal{F}$ controls the sum of squared bonuses at the observed state–action pairs. Lemma 17 uses an exponential-moment argument to transfer this bound to the sum of occupancy expectations $\sum_{k,h}\mathbb{E}_{d_h^{\tau_k}}[b_{k,h}^2]$, with failure probability at most $\delta/4$. Combining this event with the planning event above by a union bound gives probability at least $1-\delta/2$ and completes the bound on T_2. The detailed argument is given in Appendix C.3. □

These two ingredients establish the theorem as follows.

Proof of Theorem 1. By Lemma 6,

$$\mathrm{Regret}_{\omega,\alpha,\tau}(K)\le T_1+T_2.$$

Lemma 7 bounds T_1 with failure probability at most $\delta/2$, while Lemma 8 bounds T_2 with failure probability at most $\delta/2$. A union bound therefore yields the claimed result with probability at least $1-\delta$. Substituting the two bounds gives Eq. 9. □

C.2 Proof of the Reward-Learning Bound

C.2.1 Technical Lemmas

In this section, we first state the auxiliary results used to prove Lemma 7. Proofs of following lemmas are deferred to Appendix C.2.3.

We start with the following lemma showing that the reward error could be decomposed by OMD analysis into a stability term T_S and a concentration term T_C, which we will bound respectively.

Lemma 9 (Reward error decomposition)**.** *Define*

$$T_S := \sum_{k=1}^{K}\sum_{h=1}^{H}\Big[\langle \widehat{g}_{k,h}, r_{k,h}-r_{k+1,h}\rangle - \frac{\alpha\rho}{2}\|r_{k,h}-r_{k+1,h}\|^2_{k,h}\Big], \tag{17}$$

$$T_C(r) := \sum_{k=1}^{K}\Big[\ell_k(r_k)-\ell_k(r)-\widehat{\ell}_k(r_k)+\widehat{\ell}_k(r)\Big] - \frac{\alpha(1-\rho)}{2}\sum_{k=1}^{K}\sum_{h=1}^{H}\Big[\omega\langle \hat{d}^E_h,(r_{k,h}-r_h)^2\rangle+(1-\omega)\langle \hat{d}_{k,h},(r_{k,h}-r_h)^2\rangle\Big]. \tag{18}$$

Then, the reward update in Eq. 6 optimization objective ensures that, for every $r\in\mathcal{R}$,

$$\sum_{k=1}^{K}\big(\ell_k(r_k)-\ell_k(r)\big)\le T_S+T_C(r), \qquad T_1\le T_S+\sup_{r\in\mathcal{R}}T_C(r). \tag{19}$$

For the stability term, we then show with the following lemma that, it could be bounded by terms logarithmic or constant in K and the Eluder dimension of reward class $\mathcal{R}$.

Lemma 10 (Reward stability)**.** *With $0<\rho<1$, the reward update in Eq. 6 guarantees*

$$T_S\le \frac{H(1+\alpha\omega)^2}{2\alpha\rho\omega}(1+\log K)+\frac{H(1+\alpha(1-\omega))^2}{2\alpha\rho(1-\omega)}\dim_K(\mathcal{R},4/K^2) + 2H(1+\alpha(1-\omega)). \tag{20}$$

For the concentration term, we show with the following lemmas that, it could be bounded by terms logarithmic in K or of order $\tilde{\mathcal{O}}(K/N)$

Lemma 11 (Freedman's inequality). *Let $\{X_t\}_{t=1}^n$ be a martingale-difference sequence with respect to the filtration $\{\mathcal{G}_t\}_{t=0}^n$, with $|X_t| \le b$ almost surely for a deterministic $b > 0$, and $v_t := \mathbb{E}[X_t^2 \mid \mathcal{G}_{t-1}]$. For any deterministic $0 < \gamma \le 1/b$ and $z > 0$, with probability at least $1 - e^{-z}$,*

$$\sum_{t=1}^{n} X_t \le \gamma \sum_{t=1}^{n} v_t + \frac{z}{\gamma}. \tag{21}$$

Lemma 12. *With $0 < \rho < 1$, it holds with probability at least $1 - \delta/2$ that*

$$\sup_{r\in\mathcal{R}} T_C(r) \le \mathcal{O}\Bigg(\frac{H(1+\alpha(1-\omega))^2}{\alpha(1-\omega)(1-\rho)} \log \frac{4H\overline{\mathcal{N}}_{\mathcal{R}}(1/K^2)}{\delta} + \frac{KH(1+\alpha\omega)^2}{\alpha\omega(1-\rho)N} \log \frac{4H\overline{\mathcal{N}}_{\mathcal{R}}(1/N^2)^2}{\delta} \Bigg). \tag{22}$$

C.2.2 Proof of Lemma 7

With results established in Appendix C.2.1, we provide the proof of Lemma 7 in the following.

Proof. First, by Lemmas 9, we separate the reward-learning error T_1 into two terms

$$T_1 \le T_S + \sup_{r\in\mathcal{R}} T_C(r).$$

Further, with 10 and 12, we bound the two terms respectively

$$T_S \le \frac{H(1+\alpha\omega)^2}{2\alpha\rho\omega}(1+\log K) + \frac{H(1+\alpha(1-\omega))^2}{2\alpha\rho(1-\omega)} \dim_K(\mathcal{R}, 4/K^2) + 2H(1+\alpha(1-\omega))$$

$$\sup_{r\in\mathcal{R}} T_C(r) \le \mathcal{O}\Bigg(\frac{H(1+\alpha(1-\omega))^2}{\alpha(1-\omega)(1-\rho)} \log \frac{4H\overline{\mathcal{N}}_{\mathcal{R}}(1/K^2)}{\delta} + \frac{KH(1+\alpha\omega)^2}{\alpha\omega(1-\rho)N} \log \frac{4H\overline{\mathcal{N}}_{\mathcal{R}}(1/N^2)^2}{\delta} \Bigg).$$

Finally, setting $\rho = 1/2$. Then $1/(\alpha\rho) = 2/\alpha$ and $\alpha(1-\rho) = \alpha/2$. The lower-order stability term is absorbed into the learner concentration term using

$$1 + \alpha(1-\omega) \le \frac{(1+\alpha(1-\omega))^2}{\alpha(1-\omega)}, \qquad \log \frac{4H\overline{\mathcal{N}}_{\mathcal{R}}(1/K^2)}{\delta} \ge 1.$$

This gives the stated bound up to universal numerical constants. □

C.2.3 Proofs of the Technical Lemmas

Proof of Lemma 9. The seminorm induces a positive-semidefinite bilinear form $\langle\cdot,\cdot\rangle_{k,h}$. The directional derivative of the proximal objective along $(1-t)r_{k+1} + tr \in \mathcal{R}$ is nonnegative at $t = 0$. Hence

$$\langle \widehat{g}_k, r_{k+1} - r\rangle \le \alpha\rho \sum_{h=1}^{H} \langle r_{k+1,h} - r_{k,h}, r_h - r_{k+1,h}\rangle_{k,h}.$$

Rearranging and applying the equality that $2\langle x-y, z-x\rangle = \|y-z\|^2 - \|x-z\|^2 - \|x-y\|^2$ for $\langle\cdot,\cdot\rangle_{k,h}$ and $\|\cdot\|_{k,h}$ yield

$$\langle \widehat{g}_k, r_{k+1} - r\rangle \le \frac{\alpha\rho}{2} \sum_{h=1}^{H} \Big(\|r_{k,h} - r_h\|_{k,h}^2 - \|r_{k+1,h} - r_h\|_{k,h}^2 - \|r_{k+1,h} - r_{k,h}\|_{k,h}^2 \Big). \tag{23}$$

By property of quadratic function, the empirical loss satisfies

$$\sum_{k=1}^{K}\widehat{\ell}_k(r_k)-\widehat{\ell}_k(r)=\sum_{k=1}^{K}\langle\widehat{g}_k,r_k-r\rangle - \frac{\alpha}{2}\sum_{k=1}^{K}\sum_{h=1}^{H}\Big[\omega\langle\hat{d}_h^E,(r_{k,h}-r_h)^2\rangle+(1-\omega)\langle\hat{d}_{k,h},(r_{k,h}-r_h)^2\rangle\Big]. \quad (24)$$

Split the first term in RHS through $\langle\widehat{g}_k,r_k-r\rangle=\langle\widehat{g}_k,r_{k+1}-r\rangle+\langle\widehat{g}_k,r_k-r_{k+1}\rangle$ and apply the three-point inequality, we then have

$$\begin{aligned}
&\sum_{k=1}^{K}\langle\widehat{g}_k,r_k-r\rangle\\
&=\sum_{k=1}^{K}\langle\widehat{g}_k,r_k-r_{k+1}\rangle+\sum_{k=1}^{K}\langle\widehat{g}_k,r_{k+1}-r\rangle\\
&\le\sum_{k=1}^{K}\langle\widehat{g}_k,r_k-r_{k+1}\rangle+\frac{\alpha\rho}{2}\sum_{k=1}^{K}\sum_{h=1}^{H}\Big(\|r_{k,h}-r_h\|_{k,h}^2-\|r_{k+1,h}-r_h\|_{k,h}^2-\|r_{k+1,h}-r_{k,h}\|_{k,h}^2\Big)\\
&\le\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\Big[\langle\widehat{g}_{k,h},r_{k,h}-r_{k+1,h}\rangle-\frac{\alpha\rho}{2}\|r_{k,h}-r_{k+1,h}\|_{k,h}^2\Big]}_{T_S}\\
&\qquad+\frac{\alpha\rho}{2}\sum_{k=1}^{K}\sum_{h=1}^{H}\Big[\omega\langle\hat{d}_h^E,(r_{k,h}-r_h)^2\rangle+(1-\omega)\langle\hat{d}_{k,h},(r_{k,h}-r_h)^2\rangle\Big]
\end{aligned}$$

where the last step is by telescoping and the fact that $\|\cdot\|_{0,h}=0$, $\|\cdot\|_{K,h}\ge 0$ and $\|r_{k,h}-r_h\|_{k,h}^2-\|r_{k,h}-r_h\|_{k-1,h}^2=\omega\langle\hat{d}_h^E,(r_{k,h}-r_h)^2\rangle+(1-\omega)\langle\hat{d}_{k,h},(r_{k,h}-r_h)^2\rangle$.

Plugging the above inequality back into Eq. 24 gives

$$\begin{aligned}
&\sum_{k=1}^{K}\big(\widehat{\ell}_k(r_k)-\widehat{\ell}_k(r)\big)\\
&\qquad\le T_S-\frac{\alpha(1-\rho)}{2}\sum_{k=1}^{K}\sum_{h=1}^{H}\Big[\omega\langle\hat{d}_h^E,(r_{k,h}-r_h)^2\rangle+(1-\omega)\langle\hat{d}_{k,h},(r_{k,h}-r_h)^2\rangle\Big]. \quad (25)
\end{aligned}$$

Adding the difference $\sum_{k=1}^{K}\big(\ell_k(r_k)-\ell_k(r)\big)-\sum_{k=1}^{K}\big(\widehat{\ell}_k(r_k)-\widehat{\ell}_k(r)\big)$ proves Eq. 19. □

Proof of Lemma 10. Fix k,h, recall the definition of $\hat{g}_k$ and $\|\cdot\|_{k,h}$ in Eq. 7 and 8, we split the corresponding summand of T_S into two terms.

$$\begin{aligned}
&\langle\widehat{g}_{k,h},r_{k,h}-r_{k+1,h}\rangle-\frac{\alpha\rho}{2}\|r_{k,h}-r_{k+1,h}\|_{k,h}^2\\
&=\mathbb{E}_{\hat{d}_{k,h}}\left[(1+\alpha(1-\omega)r_{k,h})(r_{k,h}-r_{k+1,h})\right]-\frac{\alpha(1-\omega)\rho}{2}\sum_{j=1}^{k}\mathbb{E}_{\hat{d}_{j,h}}\left[(r_{k,h}-r_{k+1,h})^2\right]\\
&\quad+\mathbb{E}_{\hat{d}_h^E}\left[(-1+\alpha\omega r_{k,h})(r_{k,h}-r_{k+1,h})\right]-\frac{\alpha\omega\rho k}{2}\mathbb{E}_{\hat{d}_h^E}\left[(r_{k,h}-r_{k+1,h})^2\right].
\end{aligned}$$

Since $|r_{k,h}|\le 1$, the absolute values of the expert and learner gradient weights are bounded by $1+\alpha\omega$ and $1+\alpha(1-\omega)$, respectively.

For the first term in RHS. Write $v_j := (r_{k,h}(s_{j,h}, a_{j,h}) - r_{k+1,h}(s_{j,h}, a_{j,h}))^2$, then we have

$$\mathbb{E}_{\hat{d}_k}\left[(1+\alpha(1-\omega)r_{k,h})(r_{k,h}-r_{k+1,h})\right] - \frac{\alpha(1-\omega)\rho}{2}\sum_{j=1}^{k}\mathbb{E}_{\hat{d}_j}\left[(r_{k,h}-r_{k+1,h})^2\right]$$
$$\leq (1+\alpha(1-\omega))\sqrt{v_k} - \frac{(1-\omega)\alpha\rho}{2}\sum_{j=1}^{k} v_j$$

If $\sum_{j=1}^{k} v_j = 0$, the term vanishes. Otherwise,

$$(1+\alpha(1-\omega))\sqrt{v_k} - \frac{(1-\omega)\alpha\rho}{2}\sum_{j=1}^{k} v_j \leq (1+\alpha(1-\omega))\sqrt{v_k},$$
$$(1+\alpha(1-\omega))\sqrt{v_k} - \frac{(1-\omega)\alpha\rho}{2}\sum_{j=1}^{k} v_j \leq \frac{(1+\alpha(1-\omega))^2}{2\alpha\rho(1-\omega)}\frac{v_k}{\sum_{j=1}^{k} v_j}. \tag{26}$$

the second inequality is by AM-GM. If $v_k \leq 4/K^2$, use the first bound to obtain $2(1+\alpha(1-\omega))/K$. Otherwise,

$$\frac{v_k}{\sum_{j=1}^{k} v_j} \leq \min\left\{1, \frac{v_k}{4/K^2 + \sum_{j<k} v_j}\right\}$$
$$\leq \min\left\{1, D^2_{\mathcal{R}_h}\left((s_{k,h}, a_{k,h}); \mathcal{D}_{k-1,h}; 4/K^2\right)\right\}. \tag{27}$$

where the last inequality is by the definition of $D^2_{\mathcal{R}_h}$ in Definition 1.

For the second term in RHS. We have

$$\mathbb{E}_{\hat{d}^E}\left[(-1+\alpha\omega r_{k,h})(r_{k,h}-r_{k+1,h})\right] - \frac{\alpha\omega\rho k}{2}\mathbb{E}_{\hat{d}^E}\left[(r_{k,h}-r_{k+1,h})^2\right]$$
$$\leq (1+\alpha\omega)\sqrt{\mathbb{E}_{\hat{d}^E}\left[(r_{k,h}-r_{k+1,h})^2\right]} - \frac{\alpha\omega\rho k}{2}\mathbb{E}_{\hat{d}^E}\left[(r_{k,h}-r_{k+1,h})^2\right]$$
$$\leq \frac{(1+\alpha\omega)^2}{2\alpha\omega\rho k}. \tag{28}$$

where the first inequality is by Cauchy-Schwarz, the second inequality is again by AM-GM.

Finally, combining all above, summing over k, h and using $\sum_{k=1}^{K} k^{-1} \leq 1+\log K$ and Definition 1 complete the proof. □

Proof of Lemma 11. Since $e^u \leq 1+u+u^2$ for $|u| \leq 1$, we have $\mathbb{E}\left[e^{\gamma X_i} \mid \mathcal{G}_{i-1}\right] \leq 1+\gamma^2 v_i \leq e^{\gamma^2 v_i}$. Consequently, the process

$$Z_k := \exp\left(\gamma\sum_{i=1}^{k} X_i - \gamma^2\sum_{i=1}^{k} v_i\right), \qquad k = 0, \ldots, n,$$

is a nonnegative supermartingale with $Z_0 = 1$. Markov's inequality gives

$$\mathbb{P}\left(\gamma\sum_{i=1}^{n} X_i - \gamma^2\sum_{i=1}^{n} v_i > z\right) \leq e^{-z},$$

which proves the claim. □

Proof of Lemma 12. Fix a comparator $r \in \mathcal{R}$. By Eq. 18, $T_C(r)$ decomposes into two terms, namely learner sampling error and expert sampling error as follows.

$$T_C(r) = \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\left\langle d_h^{\pi_k} - \hat{d}_{k,h}, (r_{k,h} - r_h) + \frac{\alpha(1-\omega)}{2}(r_{k,h}^2 - r_h^2)\right\rangle - \sum_{k=1}^{K}\sum_{h=1}^{H}\left\langle \hat{d}_{k,h}, \frac{\alpha(1-\omega)(1-\rho)}{2}(r_{k,h} - r_h)^2\right\rangle}_{\text{learner sampling error}}$$
$$+ \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\left\langle d_h^{E} - \hat{d}_h^{E}, -(r_{k,h} - r_h) + \frac{\alpha\omega}{2}(r_{k,h}^2 - r_h^2)\right\rangle - \sum_{k=1}^{K}\sum_{h=1}^{H}\left\langle \hat{d}_h^{E}, \frac{\alpha\omega(1-\rho)}{2}(r_{k,h} - r_h)^2\right\rangle}_{\text{expert sampling error}}$$

Learner sampling error. Fix a stage h, and define

$$Z_{k,h}(r_h) := \left\langle d_h^{\pi_k} - \hat{d}_{k,h}, (r_{k,h} - r_h) + \frac{\alpha(1-\omega)}{2}(r_{k,h}^2 - r_h^2) + \frac{\alpha(1-\omega)(1-\rho)}{2}(r_{k,h} - r_h)^2\right\rangle. \tag{29}$$

Since r_h is fixed and both r_k and π_k are $\mathcal{H}_{k-1}$-measurable, it satisfies that $\mathbb{E}[Z_{k,h}(r_h) \mid \mathcal{H}_{k-1}] = 0$ and

$$|Z_{k,h}| \le 4(1 + \alpha(1-\omega))$$
$$\mathbb{E}[Z_{k,h}(r_h)^2 \mid \mathcal{H}_{k-1}] \le (1 + \alpha(1-\omega))^2 \mathbb{E}_{d_h^{\pi_k}}[(r_{k,h} - r_h)^2],$$

where the inequalities is by the fact that $0 < \rho < 1$ and $|r| \le 1, \forall r \in \mathcal{R}$.

Next, for each h, fix a $1/K^2$-cover of $\mathcal{R}_h$, whose cardinality is at most $\overline{\mathcal{N}}_{\mathcal{R}}(1/K^2)$. Applying Lemma 11 with $X_k = Z_{k,h}$, $\gamma = \frac{\alpha(1-\omega)(1-\rho)}{4(1+\alpha(1-\omega))^2} < \frac{1}{4(1+\alpha(1-\omega))}$ and $z = \log\frac{4H\overline{\mathcal{N}}_{\mathcal{R}}(1/K^2)}{\delta}$, and a union bound then give, with probability at least $1 - \delta/4$,

$$\sum_{k=1}^{K} Z_{k,h}(r_h) \le \frac{\alpha(1-\omega)(1-\rho)}{2}\sum_{k=1}^{K}\mathbb{E}_{d_h^{\pi_k}}[(r_{k,h} - r_h)^2] + \frac{4(1+\alpha(1-\omega))^2}{\alpha(1-\omega)(1-\rho)}\log\frac{4H\overline{\mathcal{N}}_{\mathcal{R}}(1/K^2)}{\delta},$$

simultaneously for all $h \in [H]$ and r_h within the cover.

Rearranging the inequality cancels the variance term in RHS. Extending from the covers and summing over h therefore yields, for all $r \in \mathcal{R}$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left\langle d_h^{\pi_k} - \hat{d}_{k,h}, (r_{k,h} - r_h) + \frac{\alpha(1-\omega)}{2}(r_{k,h}^2 - r_h^2)\right\rangle \le \sum_{k=1}^{K}\sum_{h=1}^{H}\left\langle \hat{d}_{k,h}, \frac{\alpha(1-\omega)(1-\rho)}{2}(r_{k,h} - r_h)^2\right\rangle + \frac{4H(1+\alpha(1-\omega))^2}{\alpha(1-\omega)(1-\rho)}\log\frac{4H\overline{\mathcal{N}}_{\mathcal{R}}(1/K^2)}{\delta} + \frac{2H(1+\alpha(1-\omega))}{K}.$$

The last remainder follows from the covering argument: after rearrangement, the population and empirical reward polynomials are each $(1+\alpha(1-\omega))$-Lipschitz in r_h, so approximation at scale $1/K^2$ contributes at most $2H(1+\alpha(1-\omega))/K$ after summing over k, h.

Expert sampling error. Similarly, fix h and a pair $q_h, r_h \in \mathcal{R}_h$, define

$$Y_{i,h}(q_h, r_h) := \left\langle d_h^E - \delta_{(s_h^i, a_h^i)}, -(q_h - r_h) + \frac{\alpha\omega}{2}(q_h^2 - r_h^2) + \frac{\alpha\omega(1-\rho)}{2}(q_h - r_h)^2 \right\rangle.$$

Since q_h, r_h are fixed independently of the observations inside $\mathcal{D}^E$, we have $\mathbb{E}_{d_h^E}[Y_{i,h}(q_h, r_h)] = 0$. Further, by independence of expert trajectories in $\mathcal{D}^E$ and noticing that

$$|Y_{i,h}(q_h, r_h)| \le 4(1+\alpha\omega)$$
$$\mathbb{E}_{d_h^E}[Y_{i,h}(q_h, r_h)^2] \le (1+\alpha\omega)^2 \mathbb{E}_{d_h^E}[(q_h - r_h)^2],$$

applying Lemma 11 with $X_i = Y_{i,h}$, $\gamma = \frac{\alpha\omega(1-\rho)}{4(1+\alpha\omega)^2} < \frac{1}{4(1+\alpha\omega)}$ and $z = \log \frac{4H\overline{\mathcal{N}}_{\mathcal{R}}(1/N^2)^2}{\delta}$, and a union bound over stages h and pairs of cover elements give, with probability at least $1 - \delta/4$,

$$\sum_{i=1}^{N} Y_{i,h}(q_h, r_h) \le \frac{\alpha\omega(1-\rho)}{2} N \mathbb{E}_{d_h^E}[(q_h - r_h)^2] + \frac{4(1+\alpha\omega)^2}{\alpha\omega(1-\rho)} \log \frac{4H\overline{\mathcal{N}}_{\mathcal{R}}(1/N^2)^2}{\delta}, \tag{30}$$

simultaneously for all $h \in [H]$ and q_h, r_h within the cover.

Again, rearranging terms cancels the quadratic term in RHS. Then, dividing both sides by N, extending from the covers, plugging in rewards $q_h = r_{k,h}$ while retaining the same fixed comparator r_h, and summing up over $k \in [K]$, $h \in [H]$ yield, for all $r \in \mathcal{R}$

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H} \left\langle d_h^E - \hat{d}_h^E, -(r_{k,h} - r_h) + \frac{\alpha\omega}{2}(r_{k,h}^2 - r_h^2) \right\rangle \\
&\le \sum_{k=1}^{K}\sum_{h=1}^{H} \left\langle \hat{d}_h^E, \frac{\alpha\omega(1-\rho)}{2}(r_{k,h} - r_h)^2 \right\rangle \\
&\qquad + \frac{4(1+\alpha\omega)^2 HK}{\alpha\omega(1-\rho)N} \log \frac{4H\overline{\mathcal{N}}_{\mathcal{R}}(1/N^2)^2}{\delta} + \frac{8(1+\alpha\omega)HK}{N^2}
\end{aligned}$$

The last remainder follows by approximating both rewards q_h and r_h within $1/N^2$ in sup norm. For each h and k, this changes the reward polynomials by at most $8(1+\alpha\omega)/N^2$.

Combining the bounds. Finally, intersecting the events on which the learner and expert sampling bounds hold and applying an union bound gives, with probability at least $1 - \delta/2$, for all $r \in \mathcal{R}$,

$$\begin{aligned}
T_C(r) \le \mathcal{O}\Bigg(&\frac{H(1+\alpha(1-\omega))^2}{\alpha(1-\omega)(1-\rho)} \log \frac{4H\overline{\mathcal{N}}_{\mathcal{R}}(1/K^2)}{\delta} \\
&+ \frac{KH(1+\alpha\omega)^2}{\alpha\omega(1-\rho)N} \log \frac{4H\overline{\mathcal{N}}_{\mathcal{R}}(1/N^2)^2}{\delta} \Bigg),
\end{aligned}$$

where the cover-extension remainders are absorbed using $K, N \ge 1$. Taking the supremum over $r \in \mathcal{R}$ completes the proof.

□

C.3 Proof of the Policy-Planning Error Bound

C.3.1 Technical Lemmas

In this section, we state the auxiliary results used to prove Lemma 8.

We start by defining the possible action- and state-value function classes as follows.

Definition 4. *For the regression classes $\mathcal{F}_h$, the bonus classes in Definition 3, and the fixed reference policy π^{ref}, define*

$$\mathcal{Q}_h := \left\{ \mathrm{clip}_{[-V_{\max}, V_{\max}]}(f + b) : f \in \mathcal{F}_h,\ b \in \mathcal{B}_h \right\}, \tag{31}$$

$$\mathcal{V}_h := \left\{ s \mapsto \tau \log \mathbb{E}_{a \sim \pi_h^{\mathrm{ref}}(\cdot|s)} \exp\left(\tau^{-1} Q(s,a)\right) : Q \in \mathcal{Q}_h \right\}. \tag{32}$$

Set $\mathcal{V}_{H+1} := \{0\}$. Every $Q \in \mathcal{Q}_h$ and $V \in \mathcal{V}_h$ satisfies $\|Q\|_\infty, \|V\|_\infty \le V_{\max}$. These classes describe the algorithm's optimistic estimates.

On every possible data history, the regression update gives $\widehat{f}_{k,h} \in \mathcal{F}_h$ and Definition 3 gives $b_{k,h} \in \mathcal{B}_h$. Consequently, the clipping and soft-value updates give $\widehat{Q}_{k,h} \in \mathcal{Q}_h$ and $\widehat{V}_{k,h} \in \mathcal{V}_h$, with $\widehat{V}_{k,H+1} \equiv 0$. This membership holds without conditioning on a confidence event.

We then show that the the covering numbers of these two classes could be bounded by those of $\mathcal{F}$ and $\mathcal{B}$.

Lemma 13. *For the classes in Definition 4, it satisfies that, for all $h \in [H]$ and $\kappa > 0$*

$$\mathcal{N}_{\mathcal{V}_h}(\kappa) \le \mathcal{N}_{\mathcal{Q}_h}(\kappa) \le \mathcal{N}_{\mathcal{F}_h}(\kappa/2)\mathcal{N}_{\mathcal{B}_h}(\kappa/2). \tag{33}$$

And, for step $H+1$, $\mathcal{N}_{\mathcal{V}_{H+1}}(\kappa) = 1$.

With the bound for the covering number of $\mathcal{V}_h$ in hand, we then use the following lemma to establish a pointwise regression-error bound.

Lemma 14. *Under Assumption 1, choose β as in Theorem 1, with probability at least $1 - \delta/4$, it holds that for all $k \in [K]$, $h \in [H]$, and $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\begin{aligned} &\left| \widehat{f}_{k,h}(s,a) - u_{k,h}(s,a) - [P_h \widehat{V}_{k,h+1}](s,a) \right| \\ &\quad \le \min\left\{ 4V_{\max},\ \beta D_{\mathcal{F}_h}((s,a); \mathcal{D}_{k-1,h}; \lambda) \right\} = b_{k,h}(s,a). \end{aligned} \tag{34}$$

Based on the above lemma, we then show with the following lemma that, with high probability, our action-value estimator $\hat{Q}_{k,h}$ will serve as a point-wise optimistic upper bound of the optimal action-value function $Q^{\star,\tau}_{u_k,h}$.

Lemma 15 (Optimism)**.** *With probability at least $1 - \delta/4$, for all k, h, s, a,*

$$\widehat{Q}_{k,h}(s,a) \ge Q^{\star,\tau}_{u_k,h}(s,a), \qquad \widehat{V}_{k,h}(s) \ge V^{\star,\tau}_{u_k,h}(s), \tag{35}$$

$$\widehat{Q}_{k,h}(s,a) - u_{k,h}(s,a) - [P_h \widehat{V}_{k,h+1}](s,a) \le 2b_{k,h}(s,a). \tag{36}$$

The residual on the left of Eq. 36 need not be nonnegative under global clipping.

On the confidence event of the above lemma, we follow Lemma B.1 from Zhao et al. (2025) to derive the following bounds on the policy-planning error.

Lemma 16. *With probability at least $1 - \delta/4$,*

$$T_2 \le \frac{2H^2}{\tau} \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi_k}}[b_{k,h}^2]. \tag{37}$$

Lemma 17. *For $\beta \ge 4V_{\max}$, with probability at least $1 - \delta/4$,*

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi_k}}[b_{k,h}^2] \le 2H\beta^2 \left[\dim_K(\mathcal{F}, \lambda) + \log \frac{4H}{\delta} \right]. \tag{38}$$

C.3.2 Proof of Lemma 8

With results established in Appendix C.3.1, we provide the proof of Lemma 8 in the following.

Proof of Lemma 8. By Lemma 14, the confidence event holds with probability at least $1-\delta/4$. On this event, Lemma 15 establishes optimism and the Bellman residual bound, so Lemma 16 yields the first inequality below.

The choice in β in Theorem 1 satisfies $\beta^2 \geq 64V_{\max}^2 \log 8 > 16V_{\max}^2$, so Lemma 17 applies with failure probability at most $\delta/4$. By a union bound, its conclusion and the confidence event hold simultaneously with probability at least $1-\delta/2$. On their intersection,

$$\begin{aligned} T_2 &\leq \frac{2H^2}{\tau} \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi_k}}[b_{k,h}^2] \\ &\leq \frac{4H^3\beta^2}{\tau} \left[\dim_K(\mathcal{F}, \lambda) + \log \frac{4H}{\delta} \right]. \end{aligned}$$

Since $\log(4H/\delta) \leq \log(8H/\delta)$, this proves the stated bound. □

C.3.3 Proofs of the technical lemmas

Proof of Lemma 13. Consider two bounded action-value functions Q, Q' with $\|Q - Q'\|_\infty \leq \varepsilon$. The pointwise inequality $Q \leq Q' + \varepsilon$ implies

$$\tau \log \mathbb{E}_{a\sim\pi_h^{\mathrm{ref}}(\cdot|s)} \exp(\tau^{-1}Q(s,a)) \leq \tau \log \mathbb{E}_{a\sim\pi_h^{\mathrm{ref}}(\cdot|s)} \exp(\tau^{-1}Q'(s,a)) + \varepsilon.$$

Exchanging Q and Q' shows that the soft-value map is 1-Lipschitz in the sup norm. Applying this map to a κ-cover of $\mathcal{Q}_h$ gives a κ-cover of $\mathcal{V}_h$, proving the first inequality in Eq. 33. Clipping is also 1-Lipschitz. Thus, for $f, f' \in \mathcal{F}_h$ and $b, b' \in \mathcal{B}_h$, the corresponding members $Q, Q' \in \mathcal{Q}_h$ satisfy

$$\|Q - Q'\|_\infty \leq \|f - f'\|_\infty + \|b - b'\|_\infty.$$

Taking a $\kappa/2$-cover of each input class and applying the clipped-sum map in Eq. 31 gives a κ-cover of $\mathcal{Q}_h$ whose centers lie in $\mathcal{Q}_h$. Counting these pairs proves the second inequality. The terminal case follows from $\mathcal{V}_{H+1} = \{0\}$. □

Proof of Lemma 14. Fix $h \in [H]$, $f, g \in \mathcal{F}_h$, and $V \in \mathcal{V}_{h+1}$. Only in this proof, write

$$\begin{aligned} \Delta_j &:= f(s_{j,h}, a_{j,h}) - g(s_{j,h}, a_{j,h}), \\ \xi_j &:= V(s_{j,h+1}) - [P_h V](s_{j,h}, a_{j,h}). \end{aligned}$$

Then Δ_j is $\mathcal{H}_{j,h}$-measurable and $\mathbb{E}[\xi_j \mid \mathcal{H}_{j,h}] = 0$. Since $\|V\|_\infty \leq V_{\max}$, conditional Hoeffding's lemma gives, for every $t \in \mathbb{R}$,

$$\mathbb{E}[\exp(t\xi_j) \mid \mathcal{H}_{j,h}] \leq \exp(t^2 V_{\max}^2/2). \tag{39}$$

With the fixed parameter $\theta := 1/(8V_{\max}^2)$, it follows that

$$\mathbb{E}\left[\exp\left(\theta \left[2\Delta_j \xi_j - \frac{1}{4}\Delta_j^2 \right] \right) \middle| \mathcal{H}_{j,h} \right] \leq 1. \tag{40}$$

Iterated conditional expectation across episodes and Markov's inequality therefore imply, for each fixed k and triple (f, g, V), with probability at least $1-\varepsilon$,

$$2\sum_{j<k} \Delta_j \xi_j - \frac{1}{4} \sum_{j<k} \Delta_j^2 \leq 8V_{\max}^2 \log \frac{1}{\varepsilon}. \tag{41}$$

Take covers at radius $\kappa := V_{\max}/K^2$. By Lemma 13, at each stage the number of triples of cover elements is at most

$$\begin{aligned} &\mathcal{N}_{\mathcal{F}_h}(\kappa)^2 \mathcal{N}_{\mathcal{V}_{h+1}}(\kappa) \\ &\quad \leq \mathcal{N}_{\mathcal{F}_h}(\kappa)^2 \mathcal{N}_{\mathcal{F}_{h+1}}(\kappa/2) \mathcal{N}_{\mathcal{B}_{h+1}}(\kappa/2) \\ &\quad \leq \overline{\mathcal{N}}_{\mathcal{F}\oplus\mathcal{B}}(\kappa)^2. \end{aligned} \tag{42}$$

Apply Eq. 41 to every triple, stage, and $k \in [K]$, with $\varepsilon := \delta/[4HK\overline{\mathcal{N}}_{\mathcal{F}\oplus\mathcal{B}}(\kappa)^2]$. The resulting union bound has probability at least $1-\delta/4$. To extend it beyond the covers, choose approximants

f', g', V' within κ in sup norm and let Δ'_j, ξ'_j denote the corresponding expressions. The structural bounds imply

$$|\Delta_j|, |\Delta'_j| \le 4V_{\max}, \qquad |\xi_j|, |\xi'_j| \le 2V_{\max},$$
$$|\Delta_j - \Delta'_j| \le 2\kappa, \qquad |\xi_j - \xi'_j| \le 2\kappa.$$

Consequently,

$$\left|2\Delta_j\xi_j - \frac{1}{4}\Delta_j^2 - 2\Delta'_j\xi'_j + \frac{1}{4}(\Delta'_j)^2\right| \le 28V_{\max}\kappa. \tag{43}$$

Using $k-1 \le K$ and $\log(4HK\overline{\mathcal{N}}_{\mathcal{F}\oplus\mathcal{B}}(\kappa)^2/\delta) \le 2\log(8HK\overline{\mathcal{N}}_{\mathcal{F}\oplus\mathcal{B}}(\kappa)/\delta)$, we obtain, simultaneously for all f, g, V, k, h,

$$2\sum_{j<k}\Delta_j\xi_j \le \frac{1}{4}\sum_{j<k}\Delta_j^2 + 16V_{\max}^2\log\frac{8HK\overline{\mathcal{N}}_{\mathcal{F}\oplus\mathcal{B}}(\kappa)}{\delta} + \frac{28V_{\max}^2}{K}. \tag{44}$$

Now select the data-dependent functions $f = \widehat{f}_{k,h}$, $g = u_{k,h} + P_h\widehat{V}_{k,h+1}$, and $V = \widehat{V}_{k,h+1}$. The regression update gives $f \in \mathcal{F}_h$, Bellman completeness gives $g \in \mathcal{F}_h$, and the observation following Definition 4 gives $V \in \mathcal{V}_{h+1}$. Thus all three functions are covered by the uniform event. The regression targets are $g(s_{j,h}, a_{j,h}) + \xi_j$. Comparing the exact least-squares minimizer with g gives

$$\sum_{j<k}(\Delta_j - \xi_j)^2 \le \sum_{j<k}\xi_j^2, \qquad \text{hence} \qquad \sum_{j<k}\Delta_j^2 \le 2\sum_{j<k}\Delta_j\xi_j. \tag{45}$$

Combining this with Eq. 44 yields

$$\begin{aligned}\sum_{j<k}\Delta_j^2 &\le \frac{64}{3}V_{\max}^2\log\frac{8HK\overline{\mathcal{N}}_{\mathcal{F}\oplus\mathcal{B}}(\kappa)}{\delta} + \frac{112V_{\max}^2}{3K} \\ &\le 64V_{\max}^2\log\frac{8HK\overline{\mathcal{N}}_{\mathcal{F}\oplus\mathcal{B}}(\kappa)}{\delta},\end{aligned} \tag{46}$$

where the last step uses $K \ge 1$ and a logarithm of at least $\log 8$. For $k = 1$, the empirical sum is empty and this bound holds directly. Uniformity is essential: the current reward and continuation value may depend on the same historical transitions used for fitting; they are selected only after the uniform event has been established.

By Definition 1, the pair $\widehat{f}_{k,h}, g \in \mathcal{F}_h$ satisfies, for every (s,a),

$$\begin{aligned}|\widehat{f}_{k,h}(s,a) - g(s,a)| &\le D_{\mathcal{F}_h}((s,a);\mathcal{D}_{k-1,h};\lambda)\sqrt{\lambda + \sum_{j<k}\Delta_j^2} \\ &\le \beta D_{\mathcal{F}_h}((s,a);\mathcal{D}_{k-1,h};\lambda)\,.\end{aligned}$$

Both functions have sup norm at most $2V_{\max}$, so their pointwise difference is also at most $4V_{\max}$. Taking the minimum of these two bounds proves Eq. 34. The radius in β defined in Theorem 1 moreover satisfies $\beta^2 \ge 64V_{\max}^2\log 8 > 16V_{\max}^2$, as required by the bonus cap. □

Proof of Lemma 15. Conditioned on the event defined in Lemma 14, which holds with probability at least $1 - \delta/4$, we have

$$u_{k,h} + P_h\widehat{V}_{k,h+1} \le \widehat{f}_{k,h} + b_{k,h} \le u_{k,h} + P_h\widehat{V}_{k,h+1} + 2b_{k,h}. \tag{47}$$

The bound $\|u_{k,h}\|_\infty \le V_{\max}/H$ and the soft Bellman recursion give

$$\begin{aligned}\left|Q^{\star,\tau}_{u_k,h}(s,a)\right| &\le (H-h+1)V_{\max}/H, \\ \left|V^{\star,\tau}_{u_k,h}(s)\right| &\le (H-h+1)V_{\max}/H.\end{aligned} \tag{48}$$

The value lower bound follows by choosing π^{ref}, whose KL cost is zero; the Q lower bound follows from its Bellman equation. The upper bounds use nonnegativity of KL.

Proceed by backward induction from stage $H+1$. If $\widehat{V}_{k,h+1} \geq V^{\star,\tau}_{u_k,h+1}$, then $u_{k,h} + P_h \widehat{V}_{k,h+1} \geq Q^{\star,\tau}_{u_k,h}$. The unclipped estimate is therefore optimistic by Eq. 47. The clipping interval contains the optimal value range, so clipping preserves this inequality. Monotonicity of the soft Bellman map gives optimism for $\widehat{V}_{k,h}$ as well.

The same induction gives $u_{k,h} + P_h \widehat{V}_{k,h+1} \geq Q^{\star,\tau}_{u_k,h} \geq -V_{\max}$, so lower clipping is inactive. Upper clipping can only reduce the estimate, and Eq. 47 yields Eq. 36. The backup $u_{k,h} + P_h \widehat{V}_{k,h+1}$ may exceed $V_{\max}$, so the proof does not assert that the clipped residual is nonnegative. □

Proof of Lemma 16. We condition the proof on the event defined in Lemma 15, which holds with probability at least $1 - \delta/4$. Recall from Eq. 14 that

$$T_2 = \sum_{k=1}^{K} \left[V^{\star,\tau}_{u_k,1}(s_1) - J^{\tau}_{u_k}(\pi_k) \right]$$

Fix k. Applying Lemma 5 with $r = u_k$ and $\pi = \pi_k$ gives

$$V^{\star,\tau}_{u_k,1}(s_1) - J^{\tau}_{u_k}(\pi_k) = \tau \mathbb{E}_{P,\pi_k} \left[\sum_{h=1}^{H} \mathrm{KL}\Big(\pi_{k,h}(\cdot \mid s_h) \,\|\, \pi^{\star,\tau}_{u_k,h}(\cdot \mid s_h) \Big) \right].$$

Further, fix (k,h,s), condition on the optimism event in Lemma 15 and by the definition of $\pi_{k,h}$ and $\pi^{\star,\tau}_{u_k,h}$ we know

$$\begin{aligned}
&\tau \,\mathrm{KL}\Big(\pi_{k,h}(\cdot \mid s) \,\|\, \pi^{\star,\tau}_{u_k,h}(\cdot \mid s) \Big) \\
&\quad = \mathbb{E}_{a \sim \pi_{k,h}(\cdot|s)} \left[\widehat{Q}_{k,h}(s,a) - Q^{\star,\tau}_{u_k,h}(s,a) \right] \\
&\qquad + \tau \log \mathbb{E}_{a \sim \pi_{k,h}(\cdot|s)} \left[\exp\left(- \frac{\widehat{Q}_{k,h}(s,a) - Q^{\star,\tau}_{u_k,h}(s,a)}{\tau} \right) \right] \\
&\quad \leq \frac{1}{2\tau} \mathbb{E}_{a \sim \pi_{k,h}(\cdot|s)} \left[\big(\widehat{Q}_{k,h}(s,a) - Q^{\star,\tau}_{u_k,h}(s,a) \big)^2 \right].
\end{aligned} \tag{49}$$

where the inequality is by the optimism event $\widehat{Q}_{k,h}(s,a) - Q^{\star,\tau}_{u_k,h}(s,a) \geq 0$, and the fact that $e^{-x} \leq 1 - x + x^2/2$ for $x \geq 0$, followed by $\log y \leq y - 1$.

For the error term $\widehat{Q}_{k,h}(s,a) - Q^{\star,\tau}_{u_k,h}(s,a)$, at $h = H$, optimism, the zero terminal continuation values, and Eq. 36 give

$$0 \leq \widehat{Q}_{k,H}(s,a) - Q^{\star,\tau}_{u_k,H}(s,a) \leq 2 b_{k,H}(s,a).$$

For $h < H$, we have

$$\begin{aligned}
0 &\leq \widehat{Q}_{k,h}(s,a) - Q^{\star,\tau}_{u_k,h}(s,a) \\
&\leq 2 b_{k,h}(s,a) + \left[P_h \Big(\widehat{V}_{k,h+1} - V^{\star,\tau}_{u_k,h+1} \Big) \right](s,a) \\
&\leq 2 b_{k,h}(s,a) + \mathbb{E}_{P,\pi_k} \left[\widehat{Q}_{k,h+1}(s_{h+1}, a_{h+1}) - Q^{\star,\tau}_{u_k,h+1}(s_{h+1}, a_{h+1}) \,\middle|\, s_h = s, a_h = a \right] \\
&\leq 2 \mathbb{E}_{P,\pi_k} \left[\sum_{\ell=h}^{H} b_{k,\ell}(s_\ell, a_\ell) \,\middle|\, s_h = s, a_h = a \right].
\end{aligned} \tag{50}$$

where the first step is the optimism event, the second step is by the optimal Bellman equation and Eq. 36; the third step follows from the formulas for the two policies and nonnegativity of KL divergence. The last step follows by backward induction from the terminal bound. Together with the case $h = H$, the final bound holds for every $h \in [H]$.

Finally, combining all above, we have

$$\begin{aligned} & V^{\star,\tau}_{u_k,1}(s_1) - J^{\tau}_{u_k}(\pi_k) \\ &= \tau \mathbb{E}_{P,\pi_k}\left[\sum_{h=1}^{H} \mathrm{KL}\Big(\pi_{k,h}(\cdot \mid s_h) \,\|\, \pi^{\star,\tau}_{u_k,h}(\cdot \mid s_h)\Big)\right] \\ &\le \frac{2}{\tau}\mathbb{E}_{P,\pi_k}\left[\sum_{h=1}^{H}\left(\sum_{\ell=h}^{H} b_{k,\ell}(s_\ell,a_\ell)\right)^2\right] \\ &\le \frac{2H^2}{\tau}\sum_{h=1}^{H}\mathbb{E}_{d_h^{\pi_k}}[b_{k,h}^2]. \end{aligned} \tag{51}$$

where the first inequality combines Eqs. 49 and 50 with conditional Jensen's inequality. The last inequality follows from Cauchy–Schwarz and summing over h. Summing over k gives the claimed bound on T_2.

□

Proof of Lemma 17. Fix h and, only in this proof, let $X_{k,h} := b_{k,h}(s_{k,h},a_{k,h})^2/\beta^2$. The bonus cap and $\beta \ge 4V_{\max}$ imply $X_{k,h} \in [0,1]$. Since $e^{-x} \le 1-(1-e^{-1})x$ on $[0,1]$,

$$\mathbb{E}[e^{-X_{k,h}} \mid \mathcal{H}_{k-1}] \le \exp\!\big(-(1-e^{-1})\mathbb{E}[X_{k,h} \mid \mathcal{H}_{k-1}]\big)\,. \tag{52}$$

Iterated conditional expectation and Markov's inequality, followed by a union bound over stages, give with probability at least $1-\delta/4$,

$$\begin{aligned} \sum_{k=1}^{K}\mathbb{E}[X_{k,h} \mid \mathcal{H}_{k-1}] &\le \frac{1}{1-e^{-1}}\left[\sum_{k=1}^{K} X_{k,h} + \log\frac{4H}{\delta}\right] \\ &\le 2\sum_{k=1}^{K} X_{k,h} + 2\log\frac{4H}{\delta} \end{aligned} \tag{53}$$

simultaneously for all h. By the bonus definition,

$$\frac{b_{k,h}(s,a)^2}{\beta^2} \le \min\!\big\{1, D^2_{\mathcal{F}_h}((s,a);\mathcal{D}_{k-1,h};\lambda)\big\}\,. \tag{54}$$

For every (s,a), $b_{k,h}(s,a)$ is $\mathcal{H}_{k-1}$-measurable, so $\mathbb{E}[X_{k,h} \mid \mathcal{H}_{k-1}] = \mathbb{E}_{d_h^{\pi_k}}[b_{k,h}^2]/\beta^2$. Sum Eq. 53 over h, use Definition 1, and multiply by β^2 to conclude. □

D Proof of Corollary 1

Proof of Corollary 1. Work on the event of Theorem 1. Convexity of $\mathcal{R}$ gives $\bar{r}_K \in \mathcal{R}$. Since $\mathcal{L}(\pi,\cdot)$ is concave, $\mathcal{L}(\pi,\bar{r}_K) \ge K^{-1}\sum_{k=1}^{K}\mathcal{L}(\pi,r_k)$. The dual gap is nonnegative because

$$\inf_{\pi}\mathcal{L}(\pi,\bar{r}_K) \le \mathcal{L}(\bar{\pi}_K,\bar{r}_K) \le \sup_{r\in\mathcal{R}}\mathcal{L}(\bar{\pi}_K,r).$$

Using the definition of the episode-level mixture, we therefore obtain

$$\begin{aligned} 0 \le \sup_{r\in\mathcal{R}}\mathcal{L}(\bar{\pi},r) - \inf_{\pi}\mathcal{L}(\pi,\bar{r}) &\le \frac{1}{K}\left[\sup_{r\in\mathcal{R}}\sum_{k=1}^{K}\mathcal{L}(\pi_k,r) - \inf_{\pi}\sum_{k=1}^{K}\mathcal{L}(\pi,r_k)\right] \\ &= \frac{\mathrm{Regret}_{\omega,\alpha,\tau}(K)}{K}. \end{aligned} \tag{55}$$

Since $\lambda \le V_{\max}^2$ for the β defined in Theorem 1, we can verify that

$$\beta^2 \le 65V_{\max}^2\log\frac{8HK\overline{\mathcal{N}}_{\mathcal{F}\oplus\mathcal{B}}(V_{\max}/K^2)}{\delta}.$$

Divide Eq. 9 by K and choose the integer sample counts so that the online and expert contributions are each at most $\epsilon/2$. This gives the sufficient orders displayed in Corollary 1. Precisely, its asymptotic sample requirements mean choosing counts at least sufficiently large universal multiples of the displayed expressions, including the logarithmic factors suppressed by $\widetilde{\mathcal{O}}$; an arbitrary upper bound on the counts would not imply sufficiency. Because the complexities depend on K and N, these inequalities must hold jointly at the selected counts. No small-complexity growth rate follows from boundedness alone. □