PROVABLY EFFICIENT OFFLINE-TO-ONLINE VALUE ADAPTATION WITH GENERAL FUNCTION APPROXIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study value adaptation in offline-to-online reinforcement learning under general function approximation. Starting from an imperfect offline pretrained Q-function, the learner aims to adapt it to the target environment using only a limited amount of online interaction. We first characterize the difficulty of this setting by establishing a minimax lower bound, showing that even when the pretrained Q-function is close to optimal $Q^\star$, online adaptation can be no more efficient than pure online RL on certain hard instances. On the positive side, under a novel structural condition on the offline-pretrained value functions, we propose O2O-LSVI, an adaptation algorithm with problem-dependent sample complexity that provably improves over pure online RL. Finally, we complement our theory with neural-network experiments that demonstrate the practical effectiveness of the proposed method.

1 INTRODUCTION

Reinforcement learning (RL) has long been regarded as an effective framework for solving complex sequential decision-making problems across a wide range of domains, including robotic control (Tang et al., 2025), healthcare (Coronato et al., 2020), and language model post-training (Ouyang et al., 2022). Standard online RL paradigms typically rely on extensive interactions with the environment to optimize the policy. To reduce the need for online interaction, offline RL has emerged as an important paradigm in which the algorithm learns from a dataset collected by an existing behavioral policy instead of directly accessing the environment (Fujimoto and Gu, 2021; Kostrikov et al., 2021; Kumar et al., 2020). However, the performance of the purely offline RL approaches is usually limited by the coverage and quality of the collected dataset (Gulcehre et al., 2022; Zhan et al., 2022). In particular, when data coverage is poor or when the online environment shifts from the offline environment, the learned value function can become highly inaccurate over parts of the state-action space.

To address this issue, recent empirical approaches consider the offline-to-online adaptation setting, where an (offline) pretrained RL agent can improve through a limited amount of online interaction (Yu and Zhang, 2023; Nakamoto et al., 2023; Huang et al., 2025; Ball et al., 2023; Zhou et al., 2024). On the theoretical side, prior works have analyzed online policy adaptation starting from a suboptimal initial policy (Xie et al., 2021b), as well as hybrid RL settings that combine offline data with online exploration (Li et al., 2023; Song et al., 2022) or finetune a pretrained transition kernel (Qu et al., 2024). However, current theoretical progress still leaves a significant gap from empirical implementations, which typically leverage offline-pretrained value functions or policies in a model-free manner (Nakamoto et al., 2023; Zhou et al., 2024). This gap motivates the following question:

How can we theoretically characterize the adaptation of value functions in offline-to-online RL?

To fully characterize this problem, three aspects must be addressed. First, it is necessary to understand the inherent difficulty of adapting an arbitrary offline value function through online interaction. Second, it is important to identify the structural conditions under which an algorithm can provably benefit from the offline-pretrained value function, compared with learning from scratch through purely online RL. Third, it is essential to characterize how the performance of the adaptation algorithm is impact by the potential misspecification of the offline-pretrained value function.

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.

In this paper, we answer this question by developing a model-free adaptation framework in which the algorithm starts with a pretrained state-action value function (Q-function) obtained through offline pretraining and adapts its policy using online interactions. Under this framework, we reveal the fundamental difficulty of general offline-to-online adaptation algorithms through an information-theoretic lower bound. Importantly, we also propose an offline-to-online adaptation algorithm, O2O-LSVI, which enjoys provably efficient sample complexity under a common structural condition on the pretrained state-action value function. Our contributions can be summarized as follows:

- We formulate offline-to-online adaptation as a value-function adaptation framework and establish an information-theoretic lower bound, showing that without structural conditions on the offline-pretrained value functions, adaptation can be fundamentally as hard as training from scratch using online interaction, even when the pretrained function is sufficiently close to the optimal Q-function.
- We propose a novel structural condition, called β-separable value gaps, that reflects the empirical success of offline-to-online RL algorithms. Under condition of β-separable value gaps, the offline-pretrained Q-function is close to the optimal Q-function in some regions, while exhibiting significantly larger errors in others.
- We propose an offline-to-online adaptation algorithm called O2O-LSVI for β-separable value gaps. We show that under general function approximation, O2O-LSVI is provably more efficient than purely online RL. The key innovations are the algorithmic design and new analysis technique of a novel criterion that determines whether to use optimistic estimation or instead leverage the pretrained value function. Empirically, we demonstrate that the O2O-LSVI can be efficiently adapt a various class of offline-pretrained value functions with significant performance gain.

2 Related Works

Theoretical Analysis on Offline and Online RL. There has been extensive theoretical study of both online and offline reinforcement learning. For RL with direct online interaction with the environment, prior work has established regret and sample complexity guarantees for tabular MDPs (He et al., 2021b; Zanette and Brunskill, 2019; Jaksch et al., 2010), MDPs with linear function approximation (Jin et al., 2023; He et al., 2021a; Zhang et al., 2021; 2024), and MDPs with general function approximation (Agarwal et al., 2023; Zhao et al., 2024; Wang et al., 2020). For RL with an offline dataset, existing theory primarily focuses on pessimism based methods and establishes suboptimality guarantees under various coverage conditions (Yin and Wang, 2021; Yin et al., 2022; Xie et al., 2021a; Uehara and Sun, 2021). There is also a line of work on hybrid RL in tabular MDPs, where the learner has access to both offline data, potentially collected under dynamics shift, and online interaction, with the goal of understanding when and how offline data can improve the sample efficiency of online learning (Li et al., 2023; Qu et al., 2024; Song et al., 2022; Xie et al., 2021b).

RL with General Function Approximation. A series of theoretical work has studied reinforcement learning under general function approximation. To understand the fundamental statistical limits of this setting and shed light on the behavior of practical RL algorithms, prior works have introduced a range of complexity measures for general function classes, including Bellman Rank (Jiang et al., 2017), Witness Rank (Sun et al., 2019), Eluder Dimension (Russo and Roy, 2013), Bellman Eluder Dimension (Jin et al., 2021), Decision Estimation Coefficient (DEC) (Foster et al., 2021), Admissible Bellman Characterization (Chen et al., 2022), Generalized Eluder Dimension (Agarwal et al., 2023), and Generalized Eluder Coefficient (GEC) (Zhong et al., 2022).

Offline-to-Online RL. Many prior works have investigated how to efficiently leverage offline data in online reinforcement learning, as well as how to fine-tune offline-pretrained policies through online interaction in the offline-to-online RL setting. In the first line of work, existing methods initialize the replay buffer with offline data (Ball et al., 2023; Song et al., 2022; Zhou et al., 2023; Tan et al., 2024), achieving both empirical improvements and provable theoretical guarantees. The second line of work (Nakamoto et al., 2023; Yu and Zhang, 2023; Zhou et al., 2024) study how to directly adapt or calibrate an offline-pretrained Q-function to the online environment in order to demonstrate substantial empirical gains. Our work is primarily concerned with this second setting, in which an imperfect offline-pretrained Q-function is given and the objective is to adapt it using online interactions.

3 PRELIMINARIES

We consider the setting of episodic Markov Decision Processes (MDPs) $\langle \mathcal{S}, \mathcal{A}, H, P, r \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces. We denote H as the horizon length. $P = \{P_h\}_{[H]}$ is the transition dynamics where $P_h : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ is the transition probability for the future state given current state and action. The reward is denoted by $r = \{r_h\}_{[H]}$ where $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$. We use a fixed initial state s_1; an initial-state distribution is handled by averaging. For stochastic rewards, $r_h(s, a)$ denotes the conditional mean and $R_h^k \in [0, 1]$ denotes the observed reward.

A policy of an agent interacting with this MDP is $\pi : \mathcal{S} \times [H] \to \mathcal{A}$, where $a = \pi_h(s)$ is an action the policy takes at timestep h. For each $h \in [H]$, there exists a state value function $V_h^\pi : \mathcal{S} \to \mathbb{R}$ under policy π, representing the expected future rewards from the current state s, defined as:

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}\left(s_{h'}, \pi_{h'}(s_{h'})\right) | s_h = s\right], \qquad \forall s \in \mathcal{S}, h \in [H].$$

Furthermore, we can also define the state-action value function $Q_h^\pi(s, a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ under a policy π, which depicts the expected future rewards given current state-action pairs (s, a):

$$Q_h^\pi(s, a) = r_h(s, a) + \mathbb{E}\left[\sum_{h'=h+1}^{H} r_{h'}\left(s_{h'}, \pi_{h'}(s_{h'})\right) \,\middle|\, s_h{=}s, a_h{=}a\right], \ \forall (s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H].$$

Within this formulation, the optimal policy $\pi^\star$ can obtain the optimal value function $V_h^\star(s) = \sup_\pi V_h^\pi(s)$ for every $h \in [H]$ and $s \in \mathcal{S}$. The Bellman equation is therefore written by

$$Q_h^\pi(s, a) = r_h(s, a) + [P_h V_{h+1}^\pi](s, a), \quad V_h^\pi(s) = Q_h^\pi(s, \pi_h(s)), \quad V_{H+1}^\pi(s) = 0,$$

for a policy π and for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, where we use the shorthand

$$[P_h V_{h+1}](s, a) := \mathbb{E}_{s' \sim P_h(\cdot|s,a)} V_{h+1}(s').$$

Similarly, the Bellman optimality equation is:

$$Q_h^\star(s, a) = r_h(s, a) + [P_h V_{h+1}^\star](s, a), \quad V_h^\star(s) = \max_{a \in \mathcal{A}} Q_h^\star(s, a), \quad V_{H+1}^\star(s) = 0,$$

for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$. Through the online interaction, the agent would like to minimize the pseudo regret defined by the cumulative gap between optimal value function and the value function of the output policy π_k for $k \in [K]$ rounds, i.e.,

$$\text{Regret}(K) = \sum_{k=1}^{K} \left[V_1^\star(s_1) - V_1^{\pi^k}(s_1)\right]. \tag{1}$$

RL with General Function Approximation. We consider model-free reinforcement learning with general function approximation. Specifically, we assume the Q-function in the algorithm is within the general function class $\mathcal{F}$. Additionally, we follow the definition of generalized Eluder dimension (Agarwal et al., 2023) to measure the complexity of exploration in the MDP:

Definition 3.1 (Generalized eluder dimension)**.** Let $\lambda > 0$ and $h \in [H]$, a sequence of state-action pairs $Z_h = \{z_{i,h} = (s_h^i, a_h^i)\}_{i \in [K]}$ be given. The generalized eluder dimension of a function class $\mathcal{F}_h : \mathcal{S} \times \mathcal{A} \to [0, H]$ with respect to λ is defined by $\dim_K(\mathcal{F}_h) := \sup_{|Z_h|=K} \dim(\mathcal{F}_h, Z_h)$, where:

$$\dim(\mathcal{F}_h, Z_h) := \sum_{i=1}^{K} \min\left(H^2, D_{\mathcal{F}_h}^2(z_{i,h}; z_{[i-1],h})\right),$$

$$D_{\mathcal{F}_h}^2(z_{i,h}; z_{[i-1],h}) := \sup_{f_1, f_2 \in \mathcal{F}_h} \frac{\left(f_1(z_{i,h}) - f_2(z_{i,h})\right)^2}{\sum_{l \in [i-1]} \left(f_1(z_{l,h}) - f_2(z_{l,h})\right)^2 + \lambda},$$

we denote $\dim_K(\mathcal{F}) = H^{-1} \sum_{h \in [H]} \dim_K(\mathcal{F}_h)$ for simplicity.

Remark 3.2. The generalized Eluder dimension in Definition 3.1 is the unweighted ($\sigma \equiv 1$) uncertainty-sum complexity. Its comparison with the standard Eluder dimension (Russo and Van Roy, 2013) involves scale-dependent logarithmic factors; we retain the generalized quantity directly. For simplicity of presentation and to maintain focus on offline-to-online adaptation, we do not use variance-weighted regression. Our analysis instead uses concentration for nonnegative predictable costs; pursuing sharper horizon dependence using variance-aware techniques (Zhao et al., 2024) is left to future work.

We further assume the bonus oracle used in our algorithm under the general function approximation developed in (Zhao et al., 2024):

Assumption 3.3 (Bonus oracle). *We assume that a computable function $\bar{D}^2_{\mathcal{F}}(z, z_{[k-1]})$ satisfies*

$$D^2_{\mathcal{F}}(z, z_{[k-1]}) \le \bar{D}^2_{\mathcal{F}}(z, z_{[k-1]}) \le C D^2_{\mathcal{F}}(z, z_{[k-1]}),$$

where C is an absolute constant and $D^2_{\mathcal{F}}(z, z_{[k-1]})$ is the uncertainty measurement in Definition 3.1.

The following covering numbers are useful when analyzing the general function approximation.

Definition 3.4 (Covering numbers). For any $\kappa > 0$, we define the following covering numbers:

1. For each $h \in [H]$, there exists a κ-cover $(\mathcal{F}_h)_\kappa \subseteq \mathcal{F}_h$ with cardinality $|(\mathcal{F}_h)_\kappa| \le \mathcal{N}_\kappa(\mathcal{F}_h)$ such that, for every $f \in \mathcal{F}_h$, one can find $f' \in (\mathcal{F}_h)_\kappa$ satisfying $\|f - f'\|_\infty \le \kappa$. We further define the uniform covering number of $\mathcal{F}$ at scale κ as $\mathcal{N}_\kappa(\mathcal{F}) := \max_{h\in[H]} \mathcal{N}_\kappa(\mathcal{F}_h)$.

2. Let $\mathcal{B} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denote a bonus function class such that, for any $0 \le k \le K$, $z_{[k]} \in (\mathcal{S} \times \mathcal{A})^k$, the oracle defined in Definition 3.3, $\bar{D}_{\mathcal{F}}(\cdot; z_{[k]})$, belongs to $\mathcal{B}$.

3. For the bonus class $\mathcal{B}$, there exists an κ-cover $(\mathcal{B})_\kappa \subseteq \mathcal{B}$ with size $|(\mathcal{B})_\kappa| \le \mathcal{N}_\kappa(\mathcal{B})$ such that, for every $b \in \mathcal{B}$, there exists $b' \in (\mathcal{B})_\kappa$ satisfying $\|b - b'\|_\infty \le \kappa$.

All covers used by the theoretical algorithm are finite, fixed before online interaction, and have deterministic tie-breaking. Write $(\mathcal{F}_h)_\kappa$ and $(\mathcal{B})_\kappa$ for their chosen representatives.

The following remark reduces the notation in general function approximations into linear MDPs (Jin et al., 2023) where the episode MDP with horizon H if the transition kernel and reward are

$$P_h(\cdot|s,a) = \textstyle\sum_{i=1}^d \boldsymbol{\phi}^{(i)}(s,a)\boldsymbol{\mu}_h^{(i)}(\cdot), \quad r_h(s,a) = \boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta}_h \in [0,1], \quad \boldsymbol{\phi}(s,a), \boldsymbol{\theta} \in \mathbb{R}^d$$

Remark 3.5 (Reduction to linear MDPs). In a d-dimensional linear MDP, $Q_h^\pi = \langle \mathbf{w}_h^\pi, \boldsymbol{\phi} \rangle$ (Jin et al., 2023). For the usual bounded linear classes with $\|\phi(s,a)\|_2 \le 1$ and coefficient norm at most $W = \mathcal{O}(H\sqrt{d})$, taking the scalar ridge $\lambda = H^2$ gives $\dim_K(\mathcal{F}) = \widetilde{\mathcal{O}}(d)$, $\log \mathcal{N}_\kappa(\mathcal{F}) = \widetilde{\mathcal{O}}(d)$, and $\log \mathcal{N}_\kappa(\mathcal{B}) = \widetilde{\mathcal{O}}(d^2)$ for the exact uncertainty oracle. The latter depends on the empirical feature Gram matrix; its square is bounded above by $\phi(s,a)^\top(\Sigma_{k-1,h} + \lambda I/(4W^2))^{-1}\phi(s,a)$, where $\Sigma_{k-1,h} = \sum_{i<k} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top$. The scalar-ridge uncertainty need not equal this elliptical upper bound. Appendix A supplies the covering argument; oracle computation remains subject to Assumption 3.3.

4 Offline-to-Online Value Adaptation: Setup and Hardness

We study offline-to-online value adaptation, where the agent's only input from offline pretraining is a state-action value function Q_{ref}, and adaptation proceeds through online interaction.

This framework is *pretrain-agnostic*: unlike approaches that require knowledge of the source transition kernel (Qu et al., 2024), behavioral-policy coverage (Xie et al., 2021b), or access to the offline dataset (Song et al., 2022), it does not require these details of how Q_{ref} was obtained. This abstraction accommodates practical offline-to-online RL (Yu and Zhang, 2023; Ball et al., 2023; Huang et al., 2025) and domain adaptation.

Approximation error, environment shifts, and incomplete offline coverage can each render Q_{ref} inaccurate. Examples 5.4, 5.5, and 5.6 illustrate how such structured reference errors can arise in practice. However, accuracy alone is insufficient: without additional structural assumptions, even a uniformly accurate reference may fail to improve worst-case online sample efficiency, as formalized by the following theorem.

Theorem 4.1 (Minimax lower bound). *For any horizon $H \ge 3$, dimension $d \ge 3$, tolerance $\zeta \in (0, \frac{1}{2})$, and accuracy $0 < \epsilon \le c_0 \zeta H$, where $c_0 > 0$ is an absolute constant, there exist a pretrained value function Q_{ref} and a finite class of d-dimensional linear MDPs $\mathcal{M}_\zeta$ such that*

$$|Q_{\mathrm{ref},h}(s,a) - Q_h^\star(s,a;M)| \le \zeta H, \qquad \forall h, s, a, \ M \in \mathcal{M}_\zeta. \tag{2}$$

Any algorithm given Q_{ref} requires at least $\Omega(H^3d^2/\epsilon^2)$ online episodes to achieve ϵ-suboptimal expected average regret, namely,

$$\mathbb{E}_{M,\text{Alg}}\left[\frac{1}{K}\sum_{k=1}^{K}\left(V_1^\star(s_1) - V_1^{\pi^k}(s_1)\right)\right] \le \epsilon. \tag{3}$$

The initial state is fixed, and the expectation is over the uniform distribution on this finite hard family and the learner's interaction randomness. The family may depend on ϵ; the reference is common to all its members.

The construction of the hard instance and the proof are deferred to Appendix D.1 and the following remarks would help understanding this negative result.

Remark 4.2. Theorem 4.1 yields an $\Omega(H^3d^2/\epsilon^2)$ sample complexity for the value adaptation, which can also be achieved by the purely online algorithms (He et al., 2023). This suggests that in principle it's fundamental hard to benefit from the offline pretrained state-value function in the online process, without special structural condition on the property of Q_{ref}. In addition, in the tabular case, a similar result is also obtained in Qu et al. (2024), who establish the minimax lower bound $\Omega(H^3SA/\epsilon^2)$ under dynamics-shift transfer with tabular MDPs, showing that source-environment data cannot improve worst-case target-environment sample efficiency. This result also aligns with the prior minimax lower bound for online RL (Zhou et al., 2021) for linear MDPs.

Remark 4.3. One of the related, but different setting on misspecified function approximation is studied in Du et al. (2019). In particular, Theorem 4.1 of Du et al. (2019) states that even if the Q-function misspecification level is small, online RL still requires $\Omega(2^H)$ trajectories to find the optimal policy. However, it worth highlighting that the settings of misspecification and value adaptation is fundamentally different. In particular, the ground truth optimal action-value function typically lies outside the assumed function class in the misspecified setting, so the learner faces an inherent realizability error. In the value adaptation setting, the target Q-function still lies in the function class so the algorithm may still downgrade to the purely online learning process. However, the pretrained function Q_{ref} may be biased relative to the optimal function $Q^\star$. As a result, Theorem 4.1 suggests that the this pretrained function might not bring fundamental benefits to the online RL procedure.

5 Offline-to-Online Adaptation under Separable Value Gaps

Although Theorem 4.1 shows that offline-to-online value adaptation cannot achieve improved sample efficiency in the most general setting, many practical RL tasks are far less challenging than these pathological instances. Motivated by the strong empirical performance of recent offline-to-online RL methods (Yu and Zhang, 2023; Ball et al., 2023; Huang et al., 2025), we therefore turn to a more structured setting in which the value function satisfies a *separable gap condition*. This structure allows us to leverage prior information about the discrepancy between the offline pretrained Q-function and the optimal Q-function. Under this setting, we develop an algorithm presented in Algorithm 1 for adapting an offline pretrained Q-function Q_{ref} through online interaction under general function approximation, and show that it attains provably better sample complexity than pure online RL.

5.1 Quantifying the Imperfection of Q_{ref}

To characterize the imperfection of the offline-pre-trained Q_{ref}, we introduce two additional structural conditions. We begin with a notion of separable value gaps, which captures the minimum nonzero discrepancy between the reference function and the optimal action-value function.

Definition 5.1 (β-separable value gaps)**.** Given the optimal action-value function $Q^\star$, we say that Q_{ref} is *β-separable* for some $\beta \in (0, H]$, if for any $h \in [H]$, $Q_h^\star(s,a) \neq Q_{\text{ref};h}(s,a)$, it holds that $|Q_h^\star(s,a) - Q_{\text{ref};h}(s,a)| \ge \beta$ for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$.

Intuitively, the β-separable value gaps indicates that the offline pretrained function Q_{ref} either coincides exactly with $Q^\star$ or differs from it by at least β. This structural condition reflects many empirical settings. For instance, in many RL tasks such as AntMaze in MuJoCo, different initial state might lead to distinct trajectories. As a result, the offline-pretrained Q_{ref} might estimate precisely for the covered initial states, but yields a significant estimation error for the uncovered cases.

Remark 5.2. A similar setting has been studied in Qu et al. (2024), dubbed as β-separable MDPs, where the β-separation is defined at the TV distance between the source and the target MDPs.

Our β-separable value gaps can be viewed as a "model-free" counterpart of Qu et al. (2024) and reflects more general offline-to-online cases, since the shift of the transition kernel does not naturally guarantees a shift of the optimal value function as well as the optimal policy, a β-separable MDPs might be too pessmistic to quantify the shift. In addition, the gaps in value functions reflects the fact that the offline pretraining can still generate a biased Q_{ref} because of the imperfect coverage of the behavioral policy, even with the same transition kernel.

Given the β-separable value gaps, we then quantify the amount of the state-action pairs where the pretrained Q_{ref} is inaccurate:

Assumption 5.3 (Bounded accuracy coverage)**.** *We assume the Q_{ref} satisfies the β-separable gap and for any policy π, the inaccuracy coverage of $Q_{\text{ref};h}$ is bounded by*

$$\mathbb{E}_{(s,a)\sim d_h^\pi}[\mathbb{1}(Q_{\text{ref};h}(s,a) \neq Q_h^\star(s,a))] \leq \rho, \qquad \forall h \in [H],$$

for some $\rho \in [0,1]$, where d_h^π denotes the state-action occupancy measure of policy π at stage h.

Assumption 5.3 bounds the visitation probability of inaccurate state-action pairs by ρ at each stage under every policy; it is not a table-entry fraction. In the most general case where Q_{ref} is never accurate, Assumption 5.3 holds for $\beta \to 0$ and $\rho = 1$. A similar assumption is made on in Qu et al. (2024) where only B state-action pairs are changed in the tabular MDP setting over the whole SA pairs. In the general function setting, we refer to this changing factor as ρ for the potential infinite state-action pairs.

We provide the following examples to illustrate how these errors can nevertheless satisfy the structural conditions formalized in Definition 5.1 and Assumption 5.3. We start from the first example targeting the ***misspecified function approximation*** where the reference value function Q_{ref}.

Example 5.4 (Misspecified reference function)**.** *Consider an linear MDP with misspecified reward denoted by the residual ξ on the reward as*

$$P_h(s'|s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h(s') \rangle, \quad r_h(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h \rangle + \xi_h(s,a) \in [0,1].$$

Although ξ cannot be captured by the linear MDP structure, one can still leverage feature $\phi(\cdot)$ and parameters $\mu(\cdot), \theta(\cdot)$ together with the true continuation value (in this idealized existence construction) to construct an idealized reference Q_{ref} by $Q_{\text{ref};h}(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h + \int_{\mathcal{S}} V_{h+1}^\star(s')\, \mathrm{d}\boldsymbol{\mu}_h(s') \rangle$.

It can be verified that the reference function Q_{ref} satisfies $Q_h^\star(s,a) - Q_{\text{ref};h}(s,a) = \xi_h(s,a)$. In the case that the misspecification $\xi_h(s,a) \geq \beta$ is constrained in a subset $\mathcal{S}_h' \times \mathcal{A}$, one can easily verify that this is β-separation with their coverage by ρ controlled by the visitation of subset $\mathcal{S}_h' \times \mathcal{A}$. We defer the detailed verification into Appendix F.

Then the next example describe a common ***environment shift*** where the agent suffers larger penalties for suboptimal actions at the 'critical states' in target environment.

Example 5.5 (Environment shifts)**.** *Let the source MDP denoted by $M = (P, r)$ with it's optimal Q function as $Q_{\text{ref}} = Q^{\star,M}$ and a deterministic optimal policy $\pi^\star$ with reward $r(s,a) \in [r_{\min}, 1]$ outside the zero-reward absorbing state, with $r_{\min} > 0$. Denote a set of 'critical states' $\mathcal{S}' \subset \mathcal{S}$ where the agent's suboptimal action will lead to an absorbing state s_0 with $r(s_0) = 0$ in the target MDP P' with probability η. The target kernel is*

$$P_h'(\cdot|s,a) = \begin{cases} (1-\eta)P_h(\cdot|s,a) + \eta\delta_{s_0}, & s \in \mathcal{S}', a \neq \pi^*(s), \\ P_h(\cdot|s,a), & \textit{otherwise}, \end{cases}$$

It can be verified that that for the 'critical states', the shift of Q function is controlled by

$$Q_{\text{ref};h}(s,a) - Q_h^{\star,M'}(s,a) = \eta \mathbb{1}\{s \in \mathcal{S}',\ a \neq \pi_h^\star(s)\}[P_h V_{h+1}^{\star,M}](s,a),$$

Then the next example provide a non-trivial case of missing offline coverage especially when pessimistic is applied in offline RL as common practice (Xie et al., 2021a).

Example 5.6 (Missing offline coverage)**.** *Consider an MDP with rewards in $[r_{\min}, 1]$, where $r_{\min} > 0$, and behavioral support $\mathcal{C}_h = \text{supp}(d_h^\mu)$. When the optimal action is supported in every covered state in $\mathcal{C}_h$, under infinite offline data we can build the pessimistic estimation*

$$Q_{\text{ref};h}(s,a) = \mathbb{1}[(s,a) \in \mathcal{C}_h] \cdot Q_h^\star(s,a)$$

Every uncovered pair has error at least $\beta = r_{\min}$, while $\max_a Q_{\text{ref};h}(s,a) = V_h^\star(s)$ on covered states. Remarkably, this error does not propagate through the unseen states and we do not need to assume the covered set $\mathcal{C}_h$ to be isolated in the whole state-space.

5.2 Proposed Algorithm

Below we state our proposed algorithm, O2O-LSVI, follows the least-squares value iteration (LSVI) framework (Jin et al., 2023; He et al., 2021a; Zhao et al., 2024). The algorithm consists of two main phases in each iteration. First, a reverse planning phase performs backward least-squares value updates to construct confidence bounds and update the state–action value estimates. Second, an episode-collecting phase interacts with the environment by executing the greedy policy induced by Q_h^k, thereby collecting new trajectories for subsequent updates.

5.2.1 Least-Square Regression

In O2O-LSVI, we estimate the Q-function via least-squares regression. In the original LSVI-UCB framework (Jin et al., 2023), the Q-function is estimated using least-squares regression under linear function approximation. In our setting, we extend this estimation procedure to general function approximation by leveraging a generic function class $\mathcal{F}_h$, following the paradigm developed in prior works (Wang et al., 2020; Agarwal et al., 2023; Zhao et al., 2024).

Specifically, at each iteration k and stage h, we regress toward the Bellman targets to construct upper and lower confidence approximations of $Q_h^\star$. We employ two function estimators, denoted by $\hat{f}_h^k$ and $\check{f}_h^k$, which induce the upper and lower confidence bounds $\hat{Q}_h^k$ and $\check{Q}_h^k$, respectively. The exploration bonus b_h^k is constructed using a confidence radius α_k together with the bonus oracle in Assumption 3.3. For the theoretical guarantee, regression is over the fixed cover $(\mathcal{F}_h)_\kappa$ and the oracle is rounded to $\bar{D}_{h,\kappa}^k \in (\mathcal{B})_\kappa$ within κ in sup norm; adding κ to the rounded oracle preserves an upper bonus. This finite-resolution convention controls the actual hard-switching target class without assuming the switch is continuous. We note that the proposed algorithm can be naturally extended to integrate variance-aware techniques within Agarwal et al. (2023); Zhao et al. (2024), potentially enabling tighter analyses, yet here we retain unweighted regression and a nonnegative-cost concentration argument.

5.2.2 Exploration Condition

A key design in O2O-LSVI is when to directly leverage the offline pretrained Q-function Q_{ref} and when to perform UCB-based exploration. Intuitively, if we are sufficiently confident that Q_{ref} is accurate at a state–action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, we exploit it; otherwise, we resort to UCB exploration, similar to standard online model-free RL.

Formally, for $(s, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, we trust Q_{ref} if the following inclusion holds:

$$\left[\check{Q}_h^k(s,a), \hat{Q}_h^k(s,a)\right] \subseteq \left[Q_{\mathrm{ref};h}(s,a) - \tfrac{1}{2}\beta,\ Q_{\mathrm{ref};h}(s,a) + \tfrac{1}{2}\beta\right]. \tag{4}$$

In the well-specified case, it can be proved that $Q_h^\star(s,a) \in [\check{Q}_h^k(s,a), \hat{Q}_h^k(s,a)]$. Combined with Eq.(4), this implies $Q_h^\star(s,a) \in [Q_{\mathrm{ref};h}(s,a) - \frac{1}{2}\beta,\ Q_{\mathrm{ref};h}(s,a) + \frac{1}{2}\beta]$. Under the β-separation condition, it follows that $Q_h^\star(s,a) = Q_{\mathrm{ref};h}(s,a)$ thus Q_{ref} is accurate. In the algorithm, we explicitly check (4); if it holds, we use its interval projection, and otherwise we continue to apply UCB-type exploration. Here $\Pi_{[l,u]}(q) := \min\{u, \max\{l, q\}\}$. Projection is inactive on the well-specified confidence event and safeguards the misspecified case without requiring its error level.

6 Theoretical Analysis

In this section, we present the theoretical analysis for the O2O-LSVI. We start with the Bellman completeness assumption in RL with general function approximation:

Assumption 6.1 (Bellman completeness). *Let $\mathcal{F} = \{\mathcal{F}_h\}_{h=1}^H$, where each $f_h \in \mathcal{F}_h$ maps $\mathcal{S} \times \mathcal{A}$ into $[0, H-h+1]$. For every $V : \mathcal{S} \to [0, H-h]$, there exists $f \in \mathcal{F}_h$ such that $f(s,a) = r_h(s,a) + [P_h V](s,a)$ for all (s,a). We use the common output bound $L = H$.*

Assumption 6.1 is standard assumption used in the theoretical analysis of model-free RL with general function approximation (Wang et al., 2020; Agarwal et al., 2023; Zhao et al., 2024).

Algorithm 1 O2O-LSVI

Require: Reference Q_{ref}, gap β, number of iterations K, fixed covers $(\mathcal{F}_h)_\kappa, (\mathcal{B})_\kappa$, confidence radius α_k.

Set $\kappa = (KH)^{-1}$, $\lambda = H^2$, and $H_h = H - h + 1$.
for iterations $k = 1, 2, \ldots, K$ **do**
  Set $s_1^k = s_1$ and $V_{H+1}^k = \check{V}_{H+1}^k = 0$.
  for step $h = H, H-1, \ldots, 1$ **do**
    $\hat{f}_h^k \leftarrow \mathrm{argmin}_{f \in (\mathcal{F}_h)_\kappa} \sum_{i<k} (f(s_h^i, a_h^i) - R_h^i - V_{h+1}^k(s_{h+1}^i))^2$
    $\check{f}_h^k \leftarrow \mathrm{argmin}_{f \in (\mathcal{F}_h)_\kappa} \sum_{i<k} (f(s_h^i, a_h^i) - R_h^i - \check{V}_{h+1}^k(s_{h+1}^i))^2$
    Choose $\bar{D}_{h,\kappa}^k \in (\mathcal{B})_\kappa$ with $\|\bar{D}_{h,\kappa}^k - \bar{D}_{\mathcal{F}_h}(\cdot; z_{[k-1],h})\|_\infty \le \kappa$.
    $b_h^k(s,a) \leftarrow \alpha_k(\bar{D}_{h,\kappa}^k(s,a) + \kappa)$
    $\check{Q}_h^k(s,a) \leftarrow [\check{f}_h^k(s,a) - b_h^k(s,a)]_0^{H_h}$
    $\hat{Q}_h^k(s,a) \leftarrow \max\{\check{Q}_h^k(s,a), [\hat{f}_h^k(s,a) + b_h^k(s,a)]_0^{H_h}\}$
    if $[\check{Q}_h^k(s,a), \hat{Q}_h^k(s,a)] \subseteq [Q_{\mathrm{ref};h}(s,a) - \beta/2, Q_{\mathrm{ref};h}(s,a) + \beta/2]$ **then**
      $Q_h^k(s,a) \leftarrow \Pi_{[\check{Q}_h^k(s,a), \hat{Q}_h^k(s,a)]}(Q_{\mathrm{ref};h}(s,a))$
    else
      $Q_h^k(s,a) \leftarrow \hat{Q}_h^k(s,a)$
    end if
    $V_h^k(s) \leftarrow \max_a Q_h^k(s,a)$, $\check{V}_h^k(s) \leftarrow \max_a \check{Q}_h^k(s,a)$
  end for
  for step $h = 1, 2, \ldots, H$ **do**
    Take $a_h^k \leftarrow \mathrm{argmax}_a Q_h^k(s_h^k, a)$ and observe R_h^k, s_{h+1}^k.
  end for
end for
Return the episode-level uniform mixture $\bar{\pi}$ of $\pi^1, \ldots, \pi^K$.

6.1 Well-Specified β-Separable Value Gaps

In the following theorem, we bound the regret of O2O-LSVI under well-specified β-separable value gaps in Definition 5.1 where $|Q_{\mathrm{ref},h}(s,a) - Q_h^\star(s,a)| < \beta$ directly yields $Q_{\mathrm{ref},h}(s,a) = Q_h^\star(s,a)$.

Theorem 6.2. *Under Assumptions 3.3, 6.1, and 5.3, use the fixed finite covers of Definition 3.4 in Algorithm 1. For $\delta \in (0, \frac{1}{5})$, let $\kappa = (KH)^{-1}$, $\lambda = H^2$,*

$$d_{\mathcal{F}} := \dim_K(\mathcal{F}), \quad \mathcal{N}_{\mathcal{F}} := \mathcal{N}_\kappa(\mathcal{F}), \quad \mathcal{N}_{\mathcal{B}} := \mathcal{N}_\kappa(\mathcal{B}), \quad \alpha_k = c_\alpha H \sqrt{\log(8Hk^2\mathcal{N}_{\mathcal{F}}^3 \mathcal{N}_{\mathcal{B}}/\delta)},$$

for a sufficiently large absolute constant c_α. With probability at least $1 - 5\delta$,

$$\mathrm{Regret}(K) = \widetilde{\mathcal{O}}\Big(\min\Big\{\frac{H^4 d_{\mathcal{F}} \log(\mathcal{N}_{\mathcal{F}}\mathcal{N}_{\mathcal{B}})}{\beta} + H^2\sqrt{\rho K d_{\mathcal{F}} \log(\mathcal{N}_{\mathcal{F}}\mathcal{N}_{\mathcal{B}})}, H^2\sqrt{K d_{\mathcal{F}} \log(\mathcal{N}_{\mathcal{F}}\mathcal{N}_{\mathcal{B}})}\Big\}\Big).$$

The regret result in Theorem 6.2 can be immediately translated into the following sampling complexity bound, as stated in the following corollary:

Corollary 6.3. *With the notation and assumptions of Theorem 6.2, the episode-level uniform mixture returned by Algorithm 1 is ϵ-optimal with probability at least $1 - 5\delta$ using*

$$\widetilde{\mathcal{O}}\left(\min\left\{\frac{H^4 d_{\mathcal{F}} \log(\mathcal{N}_{\mathcal{F}}\mathcal{N}_{\mathcal{B}})}{\beta\epsilon} + \frac{\rho H^4 d_{\mathcal{F}} \log(\mathcal{N}_{\mathcal{F}}\mathcal{N}_{\mathcal{B}})}{\epsilon^2}, \quad \frac{H^4 d_{\mathcal{F}} \log(\mathcal{N}_{\mathcal{F}}\mathcal{N}_{\mathcal{B}})}{\epsilon^2}\right\}\right)$$

online episodes. The mixture policy selects one learned policy uniformly at the start of an evaluation episode and follows it throughout that episode.

Several remarks are important to highlight our understanding of the performance of adaptation. First we connect our result in general function approximation to the linear MDPs.

Remark 6.4. In the d-dimensional linear MDPs (Jin et al., 2023), it can be verified that the sampling complexity of Algorithm 1 reduces to $\tilde{\mathcal{O}}(\min\{(H^4 d^3)/(\beta\epsilon) + (\rho H^4 d^3)/\epsilon^2, (H^4 d^3)/\epsilon^2\})$. In the

worst case, this yields a sampling complexity of $\tilde{\mathcal{O}}(H^4d^3/\epsilon^2)$, which matches the known result of the LSVI-UCB (Jin et al., 2023). The dependence on ϵ of the upper bound matches the minimax lower bound in Theorem 4.1. In contrast, the sampling complexity upper bound still exhibits an $\tilde{\mathcal{O}}(dH)$ gap to the minimax lower bound in Theorem 4.1. Closing this horizon-and-dimension gap may require a variance-aware extension of Algorithm 1, along the lines of Zhou et al. (2021); Zhao et al. (2024); He et al. (2023); this is not established here.

The following remark demonstrates when the O2O-LSVI can improve the sample complexity

Remark 6.5. In the linear MDP setting, the worst-case sample complexity $\widetilde{\mathcal{O}}(H^4d^3/\epsilon^2)$ matches that of LSVI-UCB (Jin et al., 2023; Wang et al., 2020). Up to logarithmic factors, the ratio of our bound to this baseline is $\min\{\epsilon/\beta + \rho, 1\}$:

- When $\epsilon \gtrsim \beta$, the online branch gives $\widetilde{\mathcal{O}}(H^4d^3/\epsilon^2)$, so the guarantee avoids an additional adaptation penalty when moderate accuracy suffices.
- When $\rho > 0$ and $\epsilon \lesssim \rho\beta$, the bound becomes $\widetilde{\mathcal{O}}(\rho H^4d^3/\epsilon^2)$. By Assumption 5.3, ρ bounds the stagewise visitation probability of the inaccurate region under every policy. A smaller ρ therefore gives a ρ-factor improvement over standard online RL in this regime.

When $\rho = 0$, the adaptation term is instead $\widetilde{\mathcal{O}}(H^4d^3/(\beta\epsilon))$.

While Remark 6.5 targets for linear MDPs, similar claims can be made to other structures thanks to the general function approximations. We extend this comparison to Qu et al. (2024) in tabular MDPs.

Remark 6.6. The result in Corollary 6.3 shares similarities with the theoretical analysis of HTRL (Qu et al., 2024), which studies policy transfer between two tabular MDPs under a β-separable dynamics difference condition, analogous to our Definition 5.1. They establish a sampling complexity of $\tilde{\mathcal{O}}\big(\min\{H^3SA/\epsilon^2, H^3|B|/\epsilon^2 + H^2S^2A/(\sigma\beta)^2\}\big)$ for HTRL and provide an interpretation similar to that in Remark 6.5. In contrast, our adaptation-overhead term scales as $\mathcal{O}(1/(\beta\epsilon))$, giving inverse-linear rather than inverse-quadratic dependence on our value-gap parameter for fixed ϵ. The gap definitions and accuracy dependence differ, so this comparison does not imply uniform dominance. Moreover, the term $|B| < SA$ in Qu et al. (2024) result characterizes the size of the region where the dynamics shift occurs. Our formulation generalizes this core idea beyond the tabular setting to continuous state–action spaces under a bounded coverage assumption.

6.2 Additional Analysis and Empirical Results

The preceding analysis assumes a well-specified β-separable reference. To account for imperfect offline pretraining, Appendix A extends our guarantees to an observed reference $\tilde{Q}_{\text{ref}}$ whose uniform deviation from a β-separable reference is at most τ. For $0 < \tau \le \beta/2$, Theorem A.2 shows that this misspecification introduces an additional $\tilde{\mathcal{O}}(HK\tau)$ term in the regret bound, without requiring the algorithm to know τ.

To complement these theoretical results, we also evaluate our method on MuJoCo AntMaze and examine its certification mechanism in a controlled linear MDP, with details provided in Appendix G. On AntMaze, the neural implementation achieves final scores comparable to or higher than those of the evaluated baselines. In the linear MDP, reference switching reduces mean final cumulative pseudo-regret by 15.6% relative to LSVI-UCB. The algorithm eventually accepts accurate reference entries while rejecting misspecified ones, illustrating the behavior of the certification mechanism alongside its finite-sample benefit.

7 Conclusion

In this paper, we studied value adaptation in offline-to-online reinforcement learning under general function approximation. We showed through a minimax lower bound that, in the worst case, adapting an imperfect pretrained Q-function can be no easier than learning from scratch. Under a novel structural condition on pretrained values, however, adaptation can be substantially more sample efficient. Motivated by this insight, we proposed O2O-LSVI, which achieves improved problem-dependent guarantees over pure online RL and performs effectively in practice. These results clarify both the limitations and the potential of offline pretraining for accelerating online RL.

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

Alekh Agarwal, Yujia Jin, and Tong Zhang. Voql: Towards optimal regret in model-free rl with nonlinear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 987–1063. PMLR, 2023.

Philip J Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. In *International Conference on Machine Learning*, pages 1577–1594. PMLR, 2023.

Zixiang Chen, Chris Junchi Li, Angela Yuan, Quanquan Gu, and Michael I Jordan. A general framework for sample-efficient function approximation in reinforcement learning. *arXiv preprint arXiv:2209.15634*, 2022.

Antonio Coronato, Muddasar Naeem, Giuseppe De Pietro, and Giovanni Paragliola. Reinforcement learning for intelligent healthcare applications: A survey. *Artificial Intelligence in Medicine*, 109: 101964, 2020. ISSN 0933-3657.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning, 2020.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021.

Caglar Gulcehre, Srivatsan Srinivasan, Jakub Sygnowski, Georg Ostrovski, Mehrdad Farajtabar, Matt Hoffman, Razvan Pascanu, and Arnaud Doucet. An empirical study of implicit regularization in deep offline rl. *arXiv preprint arXiv:2207.02099*, 2022.

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*, pages 1861–1870. Pmlr, 2018.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 4171–4180. PMLR, 2021a.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for discounted mdps. *Advances in Neural Information Processing Systems*, 34:22288–22300, 2021b.

Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for linear markov decision processes. In *International Conference on Machine Learning*, pages 12790–12822. PMLR, 2023.

Xiao Huang, Xu Liu, Enze Zhang, Tong Yu, and Shuai Li. Offline-to-online reinforcement learning with classifier-free diffusion generation. *arXiv preprint arXiv:2508.06806*, 2025.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. *Mathematics of Operations Research*, 48(3):1496–1521, 2023.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

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

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Gen Li, Wenhao Zhan, Jason D Lee, Yuejie Chi, and Yuxin Chen. Reward-agnostic fine-tuning: Provable statistical benefits of hybrid reinforcement learning. *Advances in Neural Information Processing Systems*, 36:55582–55615, 2023.

Mitsuhiko Nakamoto, Simon Zhai, Anikait Singh, Max Sobol Mark, Yi Ma, Chelsea Finn, Aviral Kumar, and Sergey Levine. Cal-ql: Calibrated offline rl pre-training for efficient online fine-tuning. *Advances in Neural Information Processing Systems*, 36:62244–62269, 2023.

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

Chengrui Qu, Laixi Shi, Kishan Panaganti, Pengcheng You, and Adam Wierman. Hybrid transfer reinforcement learning: Provable sample efficiency from shifted-dynamics data. *arXiv preprint arXiv:2411.03810*, 2024.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 2256–2264, Red Hook, NY, USA, 2013. Curran Associates Inc.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Yuda Song, Yifei Zhou, Ayush Sekhari, J Andrew Bagnell, Akshay Krishnamurthy, and Wen Sun. Hybrid rl: Using both offline and online data can make rl efficient. *arXiv preprint arXiv:2210.06718*, 2022.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pages 2898–2933. PMLR, 2019.

Kevin Tan, Wei Fan, and Yuting Wei. Hybrid reinforcement learning breaks sample size barriers in linear mdps. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 120038–120077. Curran Associates, Inc., 2024. doi: 10.52202/079017-3815.

Chen Tang, Ben Abbatematteo, Jiaheng Hu, Rohan Chandra, Roberto Martín-Martín, and Peter Stone. Deep reinforcement learning for robotics: A survey of real-world successes. *Annual Review of Control, Robotics, and Autonomous Systems*, 8(1):153–188, 2025.

Denis Tarasov, Alexander Nikulin, Dmitry Akimov, Vladislav Kurenkov, and Sergey Kolesnikov. Corl: Research-oriented deep offline reinforcement learning library. *Advances in Neural Information Processing Systems*, 36:30997–31020, 2023.

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

Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33:6123–6135, 2020.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021a.

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

Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34:4065–4078, 2021.

Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*, 2022.

Zishun Yu and Xinhua Zhang. Actor-critic alignment for offline-to-online reinforcement learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 40452–40474. PMLR, 23–29 Jul 2023.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR, 2019.

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

Weitong Zhang, Dongruo Zhou, and Quanquan Gu. Reward-free model-based reinforcement learning with linear function approximation. *Advances in Neural Information Processing Systems*, 34: 1582–1593, 2021.

Weitong Zhang, Zhiyuan Fan, Jiafan He, and Quanquan Gu. Achieving constant regret in linear markov decision processes. *Advances in Neural Information Processing Systems*, 37:130694–130738, 2024.

Heyang Zhao, Jiafan He, and Quanquan Gu. A nearly optimal and low-switching algorithm for reinforcement learning with general function approximation. *Advances in Neural Information Processing Systems*, 37:94684–94735, 2024.

Han Zhong, Wei Xiong, Sirui Zheng, Liwei Wang, Zhaoran Wang, Zhuoran Yang, and Tong Zhang. Gec: A unified framework for interactive decision making in mdp, pomdp, and beyond. *arXiv preprint arXiv:2211.01962*, 2022.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

Yifei Zhou, Ayush Sekhari, Yuda Song, and Wen Sun. Offline data enhanced on-policy policy gradient with provable guarantees. *arXiv preprint arXiv:2311.08384*, 2023.

Zhiyuan Zhou, Andy Peng, Qiyang Li, Sergey Levine, and Aviral Kumar. Efficient online reinforcement learning fine-tuning need not retain offline data. *arXiv preprint arXiv:2412.07762*, 2024.

A MISSPECIFIED β-SEPARABLE VALUE GAPS

In many practical scenarios, it may still be unrealistic to obtain a reference Q-function Q_{ref} from offline RL pretraining that exactly matches $Q^\star$ on some region, as required by Definition 5.1. Instead, the algorithm may only have access to a misspecified $\tilde{Q}_{\mathrm{ref}}$, which contains the error imperfect offline RL pretraining, as defined below

Definition A.1 (Misspecified Q_{ref})**.** O2O-LSVI does not have direct access to the true reference function Q_{ref}, but instead observes a misspecified approximation $\tilde{Q}_{\mathrm{ref}}$. We assume the misspecification error is uniformly bounded as $\eta_h(s,a) := \left|\tilde{Q}_{\mathrm{ref},h}(s,a) - Q_{\mathrm{ref},h}(s,a)\right| \le \tau, \forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$.

Based on the above definition, we can now characterize the regret upper bound of Algorithm 1 under a misspecified reference Q-function, particularly without the knowledge of misspecification level τ.

Theorem A.2. *Under the assumptions and parameter choices of Theorem 6.2, run Algorithm 1 with the observed $\tilde{Q}_{\mathrm{ref}}$ in Definition A.1, including projection on its accepted-reference branch. For $0 < \tau \le \beta/2$, with probability at least $1-5\delta$, the cumulative $\mathrm{Regret}(K)$ is bounded by*

$$\widetilde{\mathcal{O}}\Big(\min\Big\{\frac{H^4 d_{\mathcal{F}}\log(\mathcal{N}_{\mathcal{F}}\mathcal{N}_{\mathcal{B}})}{\beta} + H^2\sqrt{\rho K d_{\mathcal{F}}\log(\mathcal{N}_{\mathcal{F}}\mathcal{N}_{\mathcal{B}})}, H^2\sqrt{K d_{\mathcal{F}}\log(\mathcal{N}_{\mathcal{F}}\mathcal{N}_{\mathcal{B}})}\Big\} + HK\tau\Big).$$

The algorithm does not require knowledge of τ.

Remark A.3. Theorem A.2 shows that when the offline-pretrained Q-function is misspecified with per-step error at most τ, the regret upper bound incurs an additional $\tilde{\mathcal{O}}(HK\tau)$ term compared with the bound in the well-specified case given in Theorem 6.2. This implies that the effect of misspecification accumulates over time and can not be easily detected through the trustworthy region in Algorithm 1.

Remark A.4. As discussed in Theorem 4.1, it is fundamentally hard to improve the sample complexity using an arbitrarily misspecified pretrained function $\tilde{Q}_{\mathrm{ref}}$. For O2O-LSVI, Theorem A.2 further shows that the proposed method retains the ρ-dependent benefit in the regime of Remark 6.5, provided the requested accuracy is above the additional $\mathcal{O}(H\tau)$ misspecification floor. The bound includes an additional misspecification cost absent from the purely online guarantee; it does not by itself imply that the algorithm must perform worse. Together, these results identify separable value gaps as a sufficient structural condition for improved value transfer, not as a necessary condition for every possible transfer method.

B PROOF OF THEOREM 6.2

B.1 SETS

For the detailed proof, let $\mathcal{H}_{k,h}$ denote the history immediately before the stage-h transition in episode k, including (s_h^k, a_h^k). Let $\mathcal{H}_{k,0}$ denote the history before episode k. All planning functions in episode k are $\mathcal{H}_{k,0}$-measurable. Write

$$w_h^k(s,a) := \hat{Q}_h^k(s,a) - \check{Q}_h^k(s,a), \qquad c_h^k(s,a) := \min\{4b_h^k(s,a), H\}.$$

Let $I_h^k(s,a)$ indicate that the inclusion test (4) fails. We use the following partition of visited pairs:

$$\mathcal{M}_h := \{k \in [K] : w_h^k(s_h^k, a_h^k) > \beta/4\}, \tag{5}$$

$$\mathcal{V}_h := \{k \in [K] \setminus \mathcal{M}_h : I_h^k(s_h^k, a_h^k) = 1\}. \tag{6}$$

Thus $\mathcal{M}_h$ contains wide intervals, while $\mathcal{V}_h$ contains narrow intervals that fail the actual inclusion test. Some wide intervals may pass the test; none of the bounds below assumes that every visit to $\mathcal{M}_h$ uses the UCB branch. The threshold $\beta/4$ is only an analytical partition and does not change the algorithm's $\beta/2$ containment tolerance.

Throughout this appendix, abbreviate

$$d_h := \dim_K(\mathcal{F}_h), \quad d_{\mathcal{F}} := 1 + H^{-1}\sum_{h=1}^{H} d_h, \quad \ell := \log(16HK/\delta), \quad \alpha := \max_{k\le K}\alpha_k.$$

In particular, $\alpha = \widetilde{\mathcal{O}}(H\sqrt{\log(e\mathcal{N}_{\mathcal{F}}\mathcal{N}_{\mathcal{B}})})$ and $\alpha^2 \gtrsim H^2\ell$. Constants below may depend on the absolute oracle constant in Assumption 3.3. For sums along the realized trajectories, we occasionally suppress the arguments (s_h^k, a_h^k) in c_h^k, w_h^k, and the corresponding residuals. All pointwise comparisons retain their state–action interpretation. We will repeatedly use

$$\sum_{h=1}^{H} d_h \le H d_{\mathcal{F}}, \qquad \sum_{h=1}^{H} \sqrt{d_h} \le H\sqrt{d_{\mathcal{F}}}, \qquad \kappa HK = 1.$$

The first follows from the displayed definition of $d_{\mathcal{F}}$, and the second follows from Cauchy–Schwarz. The proof proceeds from regression confidence to optimism, then to the two nonnegative-cost bounds. The main theorem is proved by combining those bounds; proofs of the supporting lemmas are supplied afterward in the existing order.

B.2 Proof Sketch of Theorem 6.2

The first key lemma bounds the number of wide-interval visits, including terminal-stage uncertainty:

Lemma B.1 (Informal statement of Lemma B.11)**.** *On the regression-confidence event, with an additional failure probability at most δ, simultaneously for every $h \in [H]$,*

$$|\mathcal{M}_h| \lesssim \min\{K, H^2\alpha^2 d_{\mathcal{F}}/\beta^2\}.$$

The second lemma uses the coverage condition only on narrow intervals that fail certification:

Lemma B.2 (Informal statement of Lemma B.14)**.** *On the same confidence event, with an additional failure probability at most δ,*

$$\sum_{h=1}^{H} \sum_{k \in \mathcal{V}_h} c_h^k(s_h^k, a_h^k) \lesssim \alpha H\sqrt{d_{\mathcal{F}}(\rho K + \ell)} + \alpha.$$

On a certified pair, $Q_h^k = Q_h^\star$, so its positive Bellman residual vanishes. Consequently, the regret is controlled by the conditional expectation of the nonnegative trajectory cost $\sum_h I_h^k c_h^k$. Lemma E.3 converts its conditional expectations into its realized sum with an additive $\mathcal{O}(H^2\ell)$ term, rather than an unlocalized $\sqrt{K}$ term. The wide-interval contribution is $\mathcal{O}(H^2\alpha^2 d_{\mathcal{F}}/\beta)$; the remaining contribution is controlled by Lemma B.2. Independently, summing all bonuses gives the ordinary online bound. Both bounds hold on the same event and therefore can be minimized.

B.3 High Probability Events

For $\mathcal{T}_h V := r_h + P_h V$, define

$$\mathcal{E}^{\hat{f}} := \left\{ \lambda + \sum_{i<k} (\hat{f}_h^k - \mathcal{T}_h V_{h+1}^k)^2(s_h^i, a_h^i) \le \alpha_k^2, \quad \forall h, k \right\},$$

$$\mathcal{E}^{\check{f}} := \left\{ \lambda + \sum_{i<k} (\check{f}_h^k - \mathcal{T}_h \check{V}_{h+1}^k)^2(s_h^i, a_h^i) \le \alpha_k^2, \quad \forall h, k \right\}, \qquad \mathcal{E} := \mathcal{E}^{\hat{f}} \cap \mathcal{E}^{\check{f}}.$$

We establish these events for the continuation functions actually generated by Algorithm 1, including its reference-switching operation. Throughout, $\lambda = H^2$ and $\kappa = (KH)^{-1}$.

Fix an episode index k and a stage $j \in [H]$. Conditional on the supplied reference, each possible pair of regression estimates is chosen from $(\mathcal{F}_j)_\kappa^2$, and the rounded bonus oracle is chosen from $(\mathcal{B})_\kappa$. Once these three functions and the deterministic radius α_k are fixed, the clipping, ordering safeguard, inclusion test, projection, and maximization over actions determine the resulting value function uniquely. Thus, the number of possible upper continuation functions is at most $\mathcal{N}_{\mathcal{F}}^2\mathcal{N}_{\mathcal{B}}$. The lower continuation functions have no more choices, and the terminal continuation is identically zero. Let $\mathcal{W}_{k,j}$ be a deterministic family containing these functions, with $\mathcal{W}_{k,H+1} = \{0\}$. Then

$$|\mathcal{W}_{k,j}| \le 2\mathcal{N}_{\mathcal{F}}^2\mathcal{N}_{\mathcal{B}} + 1 \le 3\mathcal{N}_{\mathcal{F}}^2\mathcal{N}_{\mathcal{B}}. \tag{7}$$

All its members take values in the appropriate remaining-horizon range. The family is fixed before observing the online data because the covers, reference, and α_k are fixed. Although the algorithm selects a member using the data, we can first obtain concentration simultaneously for every member of the family. This argument does not require the inclusion test to be continuous.

Fix h, k, a target $V \in \mathcal{W}_{k,h+1}$, and a candidate $f \in (\mathcal{F}_h)_\kappa$. Write $g := \mathcal{T}_h V$ and, for $i < k$, define

$$Y_i := R_h^i + V(s_{h+1}^i), \qquad \xi_i := Y_i - g(s_h^i, a_h^i), \qquad \Delta_i := f(s_h^i, a_h^i) - g(s_h^i, a_h^i).$$

Since V and f are fixed in this step, Δ_i is $\mathcal{H}_{i,h}$-measurable, and the MDP sampling rule gives

$$\begin{aligned}\mathbb{E}[\xi_i \mid \mathcal{H}_{i,h}] &= \mathbb{E}[R_h^i + V(s_{h+1}^i) \mid \mathcal{H}_{i,h}] - g(s_h^i, a_h^i) \\ &= r_h(s_h^i, a_h^i) + [P_h V](s_h^i, a_h^i) - [\mathcal{T}_h V](s_h^i, a_h^i) = 0. \end{aligned} \tag{8}$$

Also, $Y_i, g(s_h^i, a_h^i) \in [0, H - h + 1]$, so $|\xi_i| \le H$. The conditional exponential-moment bound proved in the proof of Lemma E.1, applied to the interval $[-H, H]$, yields

$$\mathbb{E}\left[\exp\left(\frac{\xi_i \Delta_i}{2H^2} - \frac{\Delta_i^2}{8H^2}\right) \middle| \mathcal{H}_{i,h}\right] \le \exp\left(\frac{H^2}{2}\frac{\Delta_i^2}{4H^4} - \frac{\Delta_i^2}{8H^2}\right) = 1. \tag{9}$$

Iterating conditional expectations over the stage-h observations and applying Markov's inequality, we obtain, with probability at least $1 - \eta$,

$$\begin{aligned}\sum_{i<k}\left(\frac{\xi_i \Delta_i}{2H^2} - \frac{\Delta_i^2}{8H^2}\right) &\le \log(1/\eta), \\ 2\sum_{i<k} \xi_i \Delta_i &\le \frac{1}{2}\sum_{i<k}\Delta_i^2 + 4H^2 \log(1/\eta). \end{aligned} \tag{10}$$

To make the bound uniform over all targets and candidates, set

$$\eta_{k,h} := \frac{\delta}{8Hk^2 \mathcal{N}_{\mathcal{F}}^3 \mathcal{N}_{\mathcal{B}}}.$$

At each h, k, the number of pairs (f, V) is at most $3\mathcal{N}_{\mathcal{F}}^3 \mathcal{N}_{\mathcal{B}}$ by (7). Hence the total failure probability is bounded by

$$\sum_{h=1}^{H}\sum_{k=1}^{K} 3\mathcal{N}_{\mathcal{F}}^3 \mathcal{N}_{\mathcal{B}} \eta_{k,h} \le \frac{3\delta}{8}\sum_{k=1}^{\infty}\frac{1}{k^2} = \frac{\pi^2}{16}\delta < \delta. \tag{11}$$

Consequently, (10) holds simultaneously for the actual upper and lower targets and their selected regression minimizers. In particular, no concentration inequality is applied after conditioning on the selected target or on a future confidence event.

Let $\hat{f}$ be the empirical minimizer for a target V in the preceding finite family. By Assumption 6.1, $g = \mathcal{T}_h V \in \mathcal{F}_h$. Definition 3.4 therefore provides $g_\kappa \in (\mathcal{F}_h)_\kappa$ satisfying $\|g_\kappa - g\|_\infty \le \kappa$. Set $\varepsilon_i := g_\kappa(s_h^i, a_h^i) - g(s_h^i, a_h^i)$ and now let $\Delta_i := \hat{f}(s_h^i, a_h^i) - g(s_h^i, a_h^i)$. By the definition of $\hat{f}$,

$$\begin{aligned}\sum_{i<k}(\Delta_i - \xi_i)^2 &= \sum_{i<k}(\hat{f}(s_h^i, a_h^i) - Y_i)^2 \\ &\le \sum_{i<k}(g_\kappa(s_h^i, a_h^i) - Y_i)^2 = \sum_{i<k}(\varepsilon_i - \xi_i)^2. \end{aligned} \tag{12}$$

Expanding the squares, canceling $\sum_{i<k}\xi_i^2$, and using $|\varepsilon_i| \le \kappa$ gives

$$\begin{aligned}\sum_{i<k}\Delta_i^2 &\le 2\sum_{i<k}\xi_i\Delta_i + \sum_{i<k}\varepsilon_i^2 - 2\sum_{i<k}\xi_i\varepsilon_i \\ &\le 2\sum_{i<k}\xi_i\Delta_i + (k-1)\kappa^2 + 2H(k-1)\kappa \\ &\le 2\sum_{i<k}\xi_i\Delta_i + 3 \end{aligned}$$

$$\leq \frac{1}{2}\sum_{i<k}\Delta_i^2 + 4H^2\log(1/\eta_{k,h}) + 3. \tag{13}$$

The second inequality uses $|\xi_i| \leq H$; the third uses $\kappa = (KH)^{-1}$ and $k-1 \leq K$; and the last follows from the uniform version of (10). Moving the squared-error term on the right to the left yields

$$\lambda + \sum_{i<k}\Delta_i^2 \leq H^2 + 8H^2\log\left(\frac{8Hk^2\mathcal{N}_{\mathcal{F}}^3\mathcal{N}_{\mathcal{B}}}{\delta}\right) + 6 \leq \alpha_k^2, \tag{14}$$

where the last inequality holds for the sufficiently large absolute constant c_α specified in Theorem 6.2. The argument applies to both targets V_{h+1}^k and $\check{V}_{h+1}^k$. Thus it proves, in particular, the stated bound $\Pr(\mathcal{E}) \geq 1 - 2\delta$.

We also give the details behind the linear uncertainty and covering bounds. Let the coefficient set of the bounded linear class have radius $W = \mathcal{O}(H\sqrt{d})$, and suppose $\|\phi(s,a)\|_2 \leq 1$. The case $W = 0$ is immediate, so take $W > 0$. Let $\mathcal{U}$ be the set of differences of two coefficient vectors; then $\|v\|_2 \leq 2W$ and $|\langle v, \phi(s,a)\rangle| \leq H$ for $v \in \mathcal{U}$. With $\Sigma = \sum_{i<k}\phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top$, the exact uncertainty oracle satisfies

$$D^2(s,a;\Sigma) = \sup_{v\in\mathcal{U}} \frac{\langle v, \phi(s,a)\rangle^2}{\lambda + v^\top\Sigma v} \leq 1, \tag{15}$$

$$\lambda + v^\top\Sigma v \geq \frac{\lambda}{4W^2}\|v\|_2^2 + v^\top\Sigma v = v^\top\left(\Sigma + \frac{\lambda}{4W^2}I\right)v. \tag{16}$$

Here (15) uses $\lambda = H^2$ and the output bound. Writing $A = \Sigma + \lambda I/(4W^2)$, Cauchy–Schwarz gives

$$\langle v, \phi(s,a)\rangle^2 = \langle A^{1/2}v, A^{-1/2}\phi(s,a)\rangle^2 \leq (v^\top A v)(\phi(s,a)^\top A^{-1}\phi(s,a)).$$

Combining these inequalities, including the zero-vector case separately, shows that

$$D^2(s,a;\Sigma) \leq \min\left\{1, \phi(s,a)^\top\left(\Sigma + \frac{\lambda}{4W^2}I\right)^{-1}\phi(s,a)\right\}. \tag{17}$$

For a fixed stage, abbreviate $\phi_i = \phi(s_h^i, a_h^i)$, let $a_0 = \lambda/(4W^2)$, and set $A_i = a_0 I + \sum_{t=1}^i \phi_t\phi_t^\top$. For any sequence of observations, the matrix determinant identity and $\min\{1,x\} \leq 2\log(1+x)$ for $x \geq 0$ imply

$$\begin{aligned}
\sum_{i=1}^K D^2(s_h^i, a_h^i; \Sigma_{i-1}) &\leq \sum_{i=1}^K \min\{1, \phi_i^\top A_{i-1}^{-1}\phi_i\} \\
&\leq 2\sum_{i=1}^K \log(1 + \phi_i^\top A_{i-1}^{-1}\phi_i) \\
&= 2\log\frac{\det A_K}{\det A_0} \\
&\leq 2d\log\left(1 + \frac{K}{da_0}\right).
\end{aligned} \tag{18}$$

The last inequality follows from $\mathrm{tr}(A_K) \leq da_0 + K$ and the arithmetic–geometric mean inequality for the eigenvalues of A_K. Taking the supremum over sequences in Definition 3.1 gives $\dim_K(\mathcal{F}_h) = \widetilde{\mathcal{O}}(d)$.

To control the bonus covering number, let Σ, Σ' be two empirical Gram matrices and fix $v \in \mathcal{U}$. Direct subtraction gives

$$\begin{aligned}
&\left|\frac{\langle v, \phi(s,a)\rangle^2}{\lambda + v^\top\Sigma v} - \frac{\langle v, \phi(s,a)\rangle^2}{\lambda + v^\top\Sigma' v}\right| \\
&\quad = \frac{\langle v, \phi(s,a)\rangle^2 |v^\top(\Sigma' - \Sigma)v|}{(\lambda + v^\top\Sigma v)(\lambda + v^\top\Sigma' v)}
\end{aligned}$$

$$\leq \frac{H^2 \cdot 4W^2 \|\Sigma' - \Sigma\|_{\mathrm{op}}}{H^4} = \frac{4W^2}{H^2} \|\Sigma' - \Sigma\|_{\mathrm{op}}. \tag{19}$$

Using $|\sup_v a_v - \sup_v b_v| \leq \sup_v |a_v - b_v|$ and $|\sqrt{x} - \sqrt{y}| \leq \sqrt{|x-y|}$ for $x, y \geq 0$, we obtain

$$\|D(\cdot;\Sigma) - D(\cdot;\Sigma')\|_\infty \leq \frac{2W}{H} \|\Sigma - \Sigma'\|_{\mathrm{op}}^{1/2}. \tag{20}$$

The empirical Gram matrices lie in an operator-norm ball of radius K in the $d(d+1)/2$-dimensional space of symmetric matrices. A cover at radius $r = \kappa^2 H^2/(4W^2)$ therefore induces a κ-cover of the exact oracle. The usual finite-dimensional packing bound, obtained by comparing the volumes of disjoint radius-$r/2$ balls with a radius-$(K + r/2)$ ball, gives

$$\log \mathcal{N}_\kappa(\mathcal{B}) \leq \frac{d(d+1)}{2} \log\left(1 + \frac{cKW^2}{\kappa^2 H^2}\right) = \widetilde{\mathcal{O}}(d^2),$$
$$\log \mathcal{N}_\kappa(\mathcal{F}) \leq d \log\left(1 + \frac{cW}{\kappa}\right) = \widetilde{\mathcal{O}}(d),$$

for an absolute constant c. For the second line, we use $\|\langle w - w', \phi(\cdot)\rangle\|_\infty \leq \|w - w'\|_2$ and the same packing argument for the coefficient set. Representatives can be chosen within the respective sets by taking maximal separated subsets. These are the uncertainty and logarithmic covering bounds used in the linear specialization.

B.4 Optimism and Pessimism Lemmas

Lemma B.3. *On $\mathcal{E}$, for all h, k, s, a,*

$$|\hat{f}_h^k(s,a) - \mathcal{T}_h V_{h+1}^k(s,a)| \leq \alpha_k D_{\mathcal{F}_h}((s,a); z_{[k-1],h}) \leq b_h^k(s,a),$$

and the same statement holds with $\hat{f}, V$ replaced by $\check{f}, \check{V}$.

Lemma B.4. *On $\mathcal{E}$, in the well-specified setting, for all h, k, s, a,*

$$0 \leq \check{Q}_h^k \leq Q_h^\star \leq Q_h^k \leq \hat{Q}_h^k \leq H - h + 1, \qquad \check{V}_h^k \leq V_h^\star \leq V_h^k.$$

Lemma B.5 (Optimism with Q_{ref})**.** *On $\mathcal{E}$, if the inclusion test holds, then $Q_{\mathrm{ref};h}(s,a) = Q_h^\star(s,a)$ and its projection onto $[\check{Q}_h^k(s,a), \hat{Q}_h^k(s,a)]$ leaves it unchanged.*

Lemma B.6 (Optimism)**.** *On $\mathcal{E}$, $Q_h^k(s,a) \geq Q_h^\star(s,a)$ and $V_h^k(s) \geq V_h^\star(s)$ for every h, k, s, a.*

B.5 Supporting Lemmas

Lemma B.7. *For each stage h, the bonuses satisfy the deterministic bound*

$$\sum_{k=1}^K c_h^k(s_h^k, a_h^k) \leq 4\sqrt{C}\,\alpha\sqrt{Kd_h} + 8\alpha\kappa K.$$

Lemma B.8. *For every subset $M \subseteq [K]$, including data-dependent subsets,*

$$\sum_{k \in M} c_h^k(s_h^k, a_h^k) \leq 4\sqrt{C}\,\alpha\sqrt{|M|d_h} + 8\alpha\kappa|M|.$$

Lemma B.9 (Conditional regret decomposition)**.** *On $\mathcal{E}$, in the well-specified setting,*

$$\mathrm{Regret}(K) \leq \sum_{k=1}^K \mathbb{E}\left[\sum_{h=1}^H I_h^k(s_h^k, a_h^k) c_h^k(s_h^k, a_h^k) \,\middle|\, \mathcal{H}_{k,0}\right].$$

Each trajectory cost inside this expectation is nonnegative and at most H^2.

Lemma B.10 (Telescoping the confidence gap)**.** *On $\mathcal{E}$, interval projection gives $V_h^k \geq \check{V}_h^k$ and, along the greedy policy,*

$$w_h^k(s,a) \leq c_h^k(s,a) + P_h(V_{h+1}^k - \check{V}_{h+1}^k)(s,a), \qquad V_h^k(s) - \check{V}_h^k(s) \leq w_h^k(s, \pi_h^k(s)).$$

Consequently, for a fixed stage h and a selection rule with $\mathbb{1}\{k \in M\}$ measurable with respect to $\mathcal{H}_{k,h}$,

$$\sum_{k\in M} w_h^k(s_h^k, a_h^k) \le 2 \sum_{k\in M} \sum_{j=h}^{H} c_j^k(s_j^k, a_j^k) + 2H^2 \log(1/\eta)$$

on an additional event of probability at least $1-\eta$. This probability assertion is for the specified predictable selection rule, not uniformly over arbitrary retrospectively chosen subsets.

Lemma B.11 (Upper bound on $|\mathcal{M}_h|$)**.** *On $\mathcal{E}$, with an additional failure probability at most δ, simultaneously for all h,*

$$|\mathcal{M}_h| \le \min\{K, C_1 H^2 \alpha^2 d_{\mathcal{F}}/\beta^2\} \tag{21}$$

for an absolute constant C_1.

Lemma B.12. *On the events of Lemma B.11,*

$$\sum_{h=1}^{H} \sum_{k\in\mathcal{M}_h} c_h^k(s_h^k, a_h^k) \lesssim \min\{H^2\alpha^2 d_{\mathcal{F}}/\beta, \alpha H\sqrt{K d_{\mathcal{F}}}\} + \alpha.$$

Lemma B.13. *Under Assumption 5.3, on $\mathcal{E}$ and an additional event of probability at least $1-\delta$,*

$$|\mathcal{V}_h| \le 2\rho K + \log(H/\delta), \qquad h \in [H].$$

When $\rho = 0$, no inaccurate pair is visited almost surely, so $|\mathcal{V}_h| = 0$ on $\mathcal{E}$.

Lemma B.14. *On the events of Lemma B.13,*

$$\sum_{h=1}^{H} \sum_{k\in\mathcal{V}_h} c_h^k(s_h^k, a_h^k) \lesssim \alpha H\sqrt{d_{\mathcal{F}}(\rho K + \ell)} + \alpha.$$

The contribution of every certified pair to the positive Bellman residual is zero in the well-specified setting.

B.6 Detailed Proof

Proof of Theorem 6.2. We first obtain the adaptation bound and the ordinary online bound on the same high-probability event. We use the notation $d_h, d_{\mathcal{F}}, \ell$ and α introduced at the beginning of this appendix.

Define the nonnegative episode cost

$$X_k := \sum_{h=1}^{H} I_h^k(s_h^k, a_h^k) c_h^k(s_h^k, a_h^k), \qquad m_k := \mathbb{E}[X_k \mid \mathcal{H}_{k,0}].$$

The inclusion indicators lie in $\{0,1\}$ and $0 \le c_h^k \le H$, so $0 \le X_k \le H^2$, including on histories where regression confidence fails. Apply the first inequality of Lemma E.3 with $B = H^2$ and failure probability δ. It gives an event $\mathcal{G}_X$ on which

$$\sum_{k=1}^{K} m_k \le 2\sum_{k=1}^{K} X_k + 2H^2 \log(1/\delta). \tag{22}$$

Let $\mathcal{G}_M$ and $\mathcal{G}_V$ be the additional concentration events used in Lemmas B.11 and B.13, respectively. Section B.3 and those lemmas imply

$$\begin{aligned}\Pr\left((\mathcal{E}\cap\mathcal{G}_M\cap\mathcal{G}_V\cap\mathcal{G}_X)^c\right) &\le \Pr(\mathcal{E}^c) + \Pr(\mathcal{G}_M^c) + \Pr(\mathcal{G}_V^c) + \Pr(\mathcal{G}_X^c)\\ &\le 2\delta + \delta + \delta + \delta = 5\delta.\end{aligned}$$

All subsequent inequalities hold on this intersection. No independence of these events is required for the union bound.

Lemma B.9 and (22) give

$$\mathrm{Regret}(K) \overset{(a)}{\le} \sum_{k=1}^{K} m_k$$

$$\overset{(b)}{\leq} 2\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k(s_h^k,a_h^k)c_h^k(s_h^k,a_h^k) + 2H^2\ell$$
$$\overset{(c)}{\leq} 2\sum_{h=1}^{H}\sum_{k\in\mathcal{M}_h} c_h^k(s_h^k,a_h^k) + 2\sum_{h=1}^{H}\sum_{k\in\mathcal{V}_h} c_h^k(s_h^k,a_h^k) + 2H^2\ell. \tag{23}$$

Here (a) uses Lemma B.9; (b) uses (22) and $\log(1/\delta)\leq\ell$; and (c) uses the partition (5)–(6). Indeed, a failed inclusion test is either at a wide interval or at a narrow interval in $\mathcal{V}_h$. A wide interval may also pass the test, but replacing its indicator by one only increases the nonnegative upper bound.

Apply Lemma B.12 to the first sum and Lemma B.14 to the second sum in (23). For an absolute constant C_2, this gives

$$\begin{aligned}\mathrm{Regret}(K) &\leq C_2\left(\frac{H^2\alpha^2 d_{\mathcal{F}}}{\beta} + \alpha H\sqrt{d_{\mathcal{F}}(\rho K+\ell)} + \alpha + H^2\ell\right)\\ &\leq C_2\left(\frac{H^2\alpha^2 d_{\mathcal{F}}}{\beta} + \alpha H\sqrt{\rho K d_{\mathcal{F}}} + \alpha H\sqrt{d_{\mathcal{F}}\ell} + \alpha + H^2\ell\right).\end{aligned} \tag{24}$$

The second inequality uses $\sqrt{x+y}\leq\sqrt{x}+\sqrt{y}$ for $x,y\geq 0$. To account explicitly for the remaining terms, choose the absolute confidence constant large enough that $\alpha\geq H\sqrt{\ell}\geq 1$. Since $d_{\mathcal{F}}\geq 1$ and $\beta\leq H$,

$$\begin{aligned}\alpha H\sqrt{d_{\mathcal{F}}\ell} &\leq \alpha^2\sqrt{d_{\mathcal{F}}} \leq \alpha^2 d_{\mathcal{F}} \leq \frac{H^2\alpha^2 d_{\mathcal{F}}}{\beta},\\ \alpha &\leq \alpha^2 \leq \frac{H^2\alpha^2 d_{\mathcal{F}}}{\beta},\\ H^2\ell &\leq \alpha^2 \leq \frac{H^2\alpha^2 d_{\mathcal{F}}}{\beta}.\end{aligned} \tag{25}$$

Substituting these three inequalities into (24), we obtain

$$\mathrm{Regret}(K) \lesssim \frac{H^2\alpha^2 d_{\mathcal{F}}}{\beta} + \alpha H\sqrt{\rho K d_{\mathcal{F}}}. \tag{26}$$

Thus the logarithmic concentration terms have been bounded explicitly, not discarded.

Starting from (b) in (23), use $I_h^k\leq 1$ and Lemma B.7 to obtain

$$\begin{aligned}\mathrm{Regret}(K) &\leq 2\sum_{h=1}^{H}\sum_{k=1}^{K} c_h^k(s_h^k,a_h^k) + 2H^2\ell\\ &\leq 8\sqrt{C}\,\alpha\sqrt{K}\sum_{h=1}^{H}\sqrt{d_h} + 16\alpha\kappa KH + 2H^2\ell\\ &\leq 8\sqrt{C}\,\alpha H\sqrt{K d_{\mathcal{F}}} + 16\alpha + 2H^2\ell.\end{aligned} \tag{27}$$

The last inequality uses $\kappa KH = 1$ and Cauchy–Schwarz:

$$\sum_{h=1}^{H}\sqrt{d_h} \leq \sqrt{H\sum_{h=1}^{H} d_h} \leq H\sqrt{d_{\mathcal{F}}}.$$

If $K\geq\ell$, then

$$\alpha\leq\alpha H\sqrt{K d_{\mathcal{F}}}, \qquad H^2\ell\leq\alpha H\sqrt{\ell}\leq\alpha H\sqrt{K d_{\mathcal{F}}},$$

so (27) is of order $\alpha H\sqrt{K d_{\mathcal{F}}}$. If $K<\ell$, we instead use rewards in $[0,1]$ to write

$$\mathrm{Regret}(K)\leq HK\leq H\sqrt{K\ell}\leq\alpha\sqrt{K}\leq\alpha H\sqrt{K d_{\mathcal{F}}}. \tag{28}$$

Therefore, for every budget K,

$$\mathrm{Regret}(K)\lesssim\alpha H\sqrt{K d_{\mathcal{F}}}. \tag{29}$$

This argument reuses the same event $\mathcal{G}_X$; it does not require an additional concentration event or the special case $|\mathcal{M}_h| = K$.

Because (26) and (29) hold simultaneously,

$$\mathrm{Regret}(K) \lesssim \min\left\{\frac{H^2\alpha^2 d_{\mathcal{F}}}{\beta} + \alpha H\sqrt{\rho K d_{\mathcal{F}}},\quad \alpha H\sqrt{K d_{\mathcal{F}}}\right\}. \tag{30}$$

Let $\mathcal{L}_K := \log(8HK^2\mathcal{N}_{\mathcal{F}}^3\mathcal{N}_{\mathcal{B}}/\delta)$, so $\alpha = c_\alpha H\sqrt{\mathcal{L}_K}$. The individual terms become

$$\frac{H^2\alpha^2 d_{\mathcal{F}}}{\beta} = \frac{c_\alpha^2 H^4 d_{\mathcal{F}}\mathcal{L}_K}{\beta},$$
$$\alpha H\sqrt{\rho K d_{\mathcal{F}}} = c_\alpha H^2\sqrt{\rho K d_{\mathcal{F}}\mathcal{L}_K},$$
$$\alpha H\sqrt{K d_{\mathcal{F}}} = c_\alpha H^2\sqrt{K d_{\mathcal{F}}\mathcal{L}_K}.$$

Finally, $\mathcal{L}_K = \mathcal{O}(\log(8HK^2/\delta) + \log(e\mathcal{N}_{\mathcal{F}}\mathcal{N}_{\mathcal{B}}))$. Substituting into (30) gives the stated leading regret orders with the positive-complexity convention fixed in this appendix. The probability is at least $1 - 5\delta$ by the union bound established at the beginning of the proof. □

Proof of Corollary 6.3. We first identify the value of the returned mixture and then solve the two regret bounds for the required number of episodes.

Conditional on the completed training history, let J be independent and uniform on $[K]$. By definition, $\bar{\pi}$ executes π^J for an entire new episode. The law of total expectation gives

$$V_1^{\bar{\pi}}(s_1) = \mathbb{E}_J[V_1^{\pi^J}(s_1)] = \frac{1}{K}\sum_{k=1}^{K} V_1^{\pi^k}(s_1),$$

$$V_1^\star(s_1) - V_1^{\bar{\pi}}(s_1) = \frac{1}{K}\sum_{k=1}^{K}\big(V_1^\star(s_1) - V_1^{\pi^k}(s_1)\big) = \frac{\mathrm{Regret}(K)}{K}. \tag{31}$$

The mixing is across entire policies at the start of an episode, rather than an independent choice of a new policy at every stage.

To make the logarithmic dependence explicit, write $C_K := d_{\mathcal{F}}\mathcal{L}_K$, with $\mathcal{L}_K$ as in the proof of Theorem 6.2. That proof and (31) imply, for an absolute constant C_3,

$$V_1^\star(s_1) - V_1^{\bar{\pi}}(s_1) \le C_3\left(\frac{H^4 C_K}{\beta K} + H^2\sqrt{\frac{\rho C_K}{K}}\right). \tag{32}$$

It suffices to make each summand at most $\epsilon/2$. The two requirements are equivalent to

$$C_3\frac{H^4 C_K}{\beta K} \le \frac{\epsilon}{2} \iff K \ge \frac{2C_3 H^4 C_K}{\beta\epsilon},$$
$$C_3 H^2\sqrt{\frac{\rho C_K}{K}} \le \frac{\epsilon}{2} \iff K \ge \frac{4C_3^2\rho H^4 C_K}{\epsilon^2}.$$

In particular, both hold if K is at least the sum of the right-hand sides. This explains why the adaptation-overhead term contains $1/\epsilon$.

The online branch of Theorem 6.2 similarly gives

$$V_1^\star(s_1) - V_1^{\bar{\pi}}(s_1) \le C_3 H^2\sqrt{\frac{C_K}{K}} \le \epsilon \quad \text{whenever} \quad K \ge \frac{C_3^2 H^4 C_K}{\epsilon^2}.$$

The algorithm's regret satisfies both bounds on the same event. Therefore it is enough for the budget to satisfy either of these sufficient requirements, giving

$$K = \widetilde{\mathcal{O}}\left(\min\left\{\frac{H^4 d_{\mathcal{F}}\log(\mathcal{N}_{\mathcal{F}}\mathcal{N}_{\mathcal{B}})}{\beta\epsilon} + \frac{\rho H^4 d_{\mathcal{F}}\log(\mathcal{N}_{\mathcal{F}}\mathcal{N}_{\mathcal{B}})}{\epsilon^2},\quad \frac{H^4 d_{\mathcal{F}}\log(\mathcal{N}_{\mathcal{F}}\mathcal{N}_{\mathcal{B}})}{\epsilon^2}\right\}\right).$$

All complexity quantities are evaluated at the selected budget K, since the covers have resolution $(KH)^{-1}$ and the eluder quantity also depends on K. The success probability is inherited from Theorem 6.2; no additional evaluation samples or concentration inequality are needed for the value identity. This is a guarantee for the mixture and does not assert a last-iterate guarantee. □

B.7 Detailed Proof of Supporting Lemmas

Proof of Lemma B.3. Fix h, k and (s, a), and suppose that $\mathcal{E}$ holds. By Assumption 6.1, both Bellman targets $g_h^k := \mathcal{T}_h V_{h+1}^k$ and $\check{g}_h^k := \mathcal{T}_h \check{V}_{h+1}^k$ belong to $\mathcal{F}_h$. The regression minimizers also belong to $\mathcal{F}_h$, because their finite cover is a subset of this class. Thus the pair $(\hat{f}_h^k, g_h^k)$ is one of the pairs in the supremum defining $D_{\mathcal{F}_h}^2$. It follows that

$$\begin{aligned}
&|\hat{f}_h^k(s,a) - g_h^k(s,a)|^2 \\
&\quad \le D_{\mathcal{F}_h}^2((s,a); z_{[k-1],h}) \left(\lambda + \sum_{i<k} |\hat{f}_h^k(s_h^i, a_h^i) - g_h^k(s_h^i, a_h^i)|^2 \right) \\
&\quad \le \alpha_k^2 D_{\mathcal{F}_h}^2((s,a); z_{[k-1],h}). \qquad (33)
\end{aligned}$$

The first inequality uses Definition 3.1; the second uses $\mathcal{E}^{\hat{f}}$. Taking square roots and then using Assumption 3.3 and the rounding rule in Algorithm 1, we obtain

$$\begin{aligned}
|\hat{f}_h^k(s,a) - \mathcal{T}_h V_{h+1}^k(s,a)| &\le \alpha_k D_{\mathcal{F}_h}((s,a); z_{[k-1],h}) \\
&\le \alpha_k \bar{D}_{\mathcal{F}_h}((s,a); z_{[k-1],h}) \\
&\le \alpha_k (\bar{D}_{h,\kappa}^k(s,a) + \kappa) = b_h^k(s,a). \qquad (34)
\end{aligned}$$

For the lower regression, use the pair $(\check{f}_h^k, \check{g}_h^k)$ and $\mathcal{E}^{\check{f}}$ in exactly the same definition:

$$\begin{aligned}
&|\check{f}_h^k(s,a) - \check{g}_h^k(s,a)|^2 \\
&\quad \le D_{\mathcal{F}_h}^2((s,a); z_{[k-1],h}) \left(\lambda + \sum_{i<k} |\check{f}_h^k(s_h^i, a_h^i) - \check{g}_h^k(s_h^i, a_h^i)|^2 \right) \\
&\quad \le \alpha_k^2 D_{\mathcal{F}_h}^2((s,a); z_{[k-1],h}).
\end{aligned}$$

Taking square roots and repeating the oracle and rounding inequalities in (34) proves the lower statement.

We record a useful consequence of these bounds. Projection in Algorithm 1 ensures $\check{Q}_h^k \le Q_h^k \le \hat{Q}_h^k$ on every history, and hence $\check{V}_h^k \le V_h^k$. Both Bellman targets belong to $[0, H_h]$, where $H_h = H - h + 1$. The just-proved error bounds imply

$$\begin{aligned}
&\check{f}_h^k - b_h^k \le \check{g}_h^k \le g_h^k \le \hat{f}_h^k + b_h^k, \\
&\check{f}_h^k - b_h^k \ge \check{g}_h^k - 2b_h^k, \qquad \hat{f}_h^k + b_h^k \le g_h^k + 2b_h^k.
\end{aligned}$$

Clipping to $[0, H_h]$ preserves the first line's comparisons with the Bellman targets. In particular, the clipped raw upper estimate is at least the lower estimate, so the extra ordering maximum is inactive on $\mathcal{E}$. Consequently, for any fixed supplied reference,

$$\check{g}_h^k - 2b_h^k \le \check{Q}_h^k \le \check{g}_h^k \le g_h^k \le \hat{Q}_h^k \le g_h^k + 2b_h^k. \qquad (35)$$

For the two outside bounds, the lower clipping can only increase $\check{f}_h^k - b_h^k \le H_h$, and the upper clipping can only decrease $\hat{f}_h^k + b_h^k \ge 0$. This consequence uses the ordering of the planning values, not optimism relative to $Q^\star$, and will also be used in the misspecified analysis. □

Joint proof of Lemmas B.4–B.6. We prove confidence, certification, and optimism together by backward induction. This avoids using a stage-h certification statement before establishing the corresponding confidence bounds. Fix an episode k and work on $\mathcal{E}$.

At stage $H+1$, Algorithm 1 sets $V_{H+1}^k = \check{V}_{H+1}^k = 0$, and the Bellman terminal condition gives $V_{H+1}^\star = 0$. Thus $\check{V}_{H+1}^k \le V_{H+1}^\star \le V_{H+1}^k$.

Suppose inductively that, at stage $h+1$, for every state s',

$$\check{V}_{h+1}^k(s') \le V_{h+1}^\star(s') \le V_{h+1}^k(s'). \qquad (36)$$

We establish all the stage-h claims at an arbitrary pair (s, a).

First, for the upper estimator, Lemma B.3 yields

$$\begin{aligned}\hat{f}_h^k(s,a) + b_h^k(s,a) - Q_h^\star(s,a) &\geq r_h(s,a) + [P_h V_{h+1}^k](s,a) - b_h^k(s,a) + b_h^k(s,a) - Q_h^\star(s,a) \\ &= [P_h(V_{h+1}^k - V_{h+1}^\star)](s,a) \geq 0. \end{aligned} \tag{37}$$

The equality uses the Bellman optimality equation, and the final inequality uses (36) and positivity of the transition operator. Since $Q_h^\star(s,a) \in [0, H_h]$, clipping does not reverse the comparison, and the ordering maximum can only increase the upper estimate. Thus $\hat{Q}_h^k(s,a) \geq Q_h^\star(s,a)$.

Similarly, for the lower estimator,

$$\begin{aligned}\check{f}_h^k(s,a) - b_h^k(s,a) - Q_h^\star(s,a) &\leq r_h(s,a) + [P_h \check{V}_{h+1}^k](s,a) + b_h^k(s,a) - b_h^k(s,a) - Q_h^\star(s,a) \\ &= [P_h(\check{V}_{h+1}^k - V_{h+1}^\star)](s,a) \leq 0. \end{aligned} \tag{38}$$

Here we again use Lemma B.3, the Bellman optimality equation, and the induction hypothesis, respectively. Clipping preserves this comparison because $Q_h^\star(s,a) \in [0, H_h]$. We have therefore proved

$$0 \leq \check{Q}_h^k(s,a) \leq Q_h^\star(s,a) \leq \hat{Q}_h^k(s,a) \leq H_h. \tag{39}$$

Next, consider the two branches of the reference test. If the test accepts, (39) and the inclusion condition imply

$$\begin{aligned}Q_h^\star(s,a) &\in [\check{Q}_h^k(s,a), \hat{Q}_h^k(s,a)] \\ &\subseteq [Q_{\mathrm{ref};h}(s,a) - \beta/2, Q_{\mathrm{ref};h}(s,a) + \beta/2].\end{aligned}$$

Consequently, $|Q_{\mathrm{ref};h}(s,a) - Q_h^\star(s,a)| \leq \beta/2 < \beta$. Definition 5.1 excludes any nonzero discrepancy of this magnitude, so $Q_{\mathrm{ref};h}(s,a) = Q_h^\star(s,a)$. This common value lies inside the confidence interval; its projection is unchanged, giving $Q_h^k(s,a) = Q_h^\star(s,a)$. If the test rejects, the algorithm uses $Q_h^k(s,a) = \hat{Q}_h^k(s,a) \geq Q_h^\star(s,a)$. Together, the two cases give

$$0 \leq \check{Q}_h^k(s,a) \leq Q_h^\star(s,a) \leq Q_h^k(s,a) \leq \hat{Q}_h^k(s,a) \leq H_h.$$

Finally, maximizing the pointwise inequalities over actions yields

$$\begin{aligned}\check{V}_h^k(s) &= \max_a \check{Q}_h^k(s,a) \leq \max_a Q_h^\star(s,a) = V_h^\star(s), \\ V_h^k(s) &= \max_a Q_h^k(s,a) \geq \max_a Q_h^\star(s,a) = V_h^\star(s).\end{aligned}$$

This closes the backward induction. The full chain proves Lemma B.4, the accepted-reference case proves Lemma B.5, and the two-branch optimism proves Lemma B.6. No additional probability loss is incurred beyond $\mathcal{E}$. □

Proof of Lemmas B.7 and B.8. Fix a stage h and an arbitrary subset $M \subseteq [K]$. For the present calculation, write $D_k := D_{\mathcal{F}_h}((s_h^k, a_h^k); z_{[k-1],h})$. The rounding rule and Assumption 3.3 give

$$\begin{aligned}b_h^k(s_h^k, a_h^k) &= \alpha_k(\bar{D}_{h,\kappa}^k(s_h^k, a_h^k) + \kappa) \\ &\leq \alpha_k\big(\bar{D}_{\mathcal{F}_h}((s_h^k, a_h^k); z_{[k-1],h}) + 2\kappa\big) \\ &\leq \alpha(\sqrt{C}\, D_k + 2\kappa).\end{aligned} \tag{40}$$

Also, the output bound $0 \leq f \leq H$ and $\lambda = H^2$ imply

$$D_k^2 = \sup_{f_1, f_2 \in \mathcal{F}_h} \frac{|f_1(s_h^k, a_h^k) - f_2(s_h^k, a_h^k)|^2}{H^2 + \sum_{i<k} |f_1(s_h^i, a_h^i) - f_2(s_h^i, a_h^i)|^2} \leq 1. \tag{41}$$

Since $H \geq 1$, the truncation in Definition 3.1 therefore leaves these uncertainty terms unchanged. In particular,

$$\sum_{k \in M} D_k^2 \leq \sum_{k=1}^K D_k^2 = \sum_{k=1}^K \min\{H^2, D_k^2\} \leq \dim_K(\mathcal{F}_h) = d_h. \tag{42}$$

It follows that

$$\begin{aligned}\sum_{k\in M} c_h^k(s_h^k,a_h^k) &\overset{(a)}{\le} 4\sum_{k\in M} b_h^k(s_h^k,a_h^k)\\ &\overset{(b)}{\le} 4\sqrt{C}\,\alpha\sum_{k\in M} D_k + 8\alpha\kappa|M|\\ &\overset{(c)}{\le} 4\sqrt{C}\,\alpha\sqrt{|M|\sum_{k\in M} D_k^2} + 8\alpha\kappa|M|\\ &\overset{(d)}{\le} 4\sqrt{C}\,\alpha\sqrt{|M|d_h} + 8\alpha\kappa|M|. \end{aligned} \tag{43}$$

Here (a) uses $c_h^k = \min\{4b_h^k, H\}$; (b) uses (40); (c) is Cauchy–Schwarz; and (d) uses (42). If $M = \varnothing$, all these sums are zero. Every step is deterministic for a realized sequence of observations, so M may be data-dependent. This proves Lemma B.8. Taking $M = [K]$ proves Lemma B.7. □

Proof of Lemma B.9. Fix an episode k. Its planning functions and greedy policy are fixed conditional on $\mathcal{H}_{k,0}$. Define

$$e_h^k(s,a) := Q_h^k(s,a) - r_h(s,a) - [P_hV_{h+1}^k](s,a), \qquad \Delta_h^k(s) := V_h^k(s) - V_h^{\pi^k}(s).$$

For $a = \pi_h^k(s)$, the greedy choice and the policy Bellman equation give

$$\begin{aligned}\Delta_h^k(s) &= Q_h^k(s,a) - Q_h^{\pi^k}(s,a)\\ &= Q_h^k(s,a) - r_h(s,a) - [P_hV_{h+1}^{\pi^k}](s,a)\\ &= Q_h^k(s,a) - r_h(s,a) - [P_hV_{h+1}^k](s,a) + [P_h(V_{h+1}^k - V_{h+1}^{\pi^k})](s,a)\\ &= e_h^k(s,a) + [P_h\Delta_{h+1}^k](s,a). \end{aligned} \tag{44}$$

Evaluate this identity along the episode and take conditional expectations. The tower property yields

$$\mathbb{E}[\Delta_h^k(s_h^k) \mid \mathcal{H}_{k,0}] = \mathbb{E}[e_h^k(s_h^k,a_h^k) \mid \mathcal{H}_{k,0}] + \mathbb{E}[\Delta_{h+1}^k(s_{h+1}^k) \mid \mathcal{H}_{k,0}].$$

Summing this equality over $h = 1,\dots,H$ cancels the intermediate value differences. Since $s_1^k = s_1$ and $\Delta_{H+1}^k = 0$, we obtain

$$V_1^k(s_1) - V_1^{\pi^k}(s_1) = \mathbb{E}\left[\sum_{h=1}^H e_h^k(s_h^k,a_h^k)\,\middle|\,\mathcal{H}_{k,0}\right]. \tag{45}$$

This identity is algebraic and remains valid without optimism.

We next bound the residual on $\mathcal{E}$ in the well-specified case. If the inclusion test fails, $Q_h^k = \hat{Q}_h^k$, and Lemma B.3, specifically (35), gives

$$e_h^k \le 2b_h^k.$$

Moreover, $Q_h^k \le H_h \le H$ and $r_h + P_hV_{h+1}^k \ge 0$, so $e_h^k \le H$. Hence $e_h^k \le \min\{2b_h^k, H\} \le c_h^k$ on this branch. If the test accepts, Lemma B.5 gives $Q_h^k = Q_h^\star$, and Lemma B.6 implies

$$\begin{aligned}e_h^k(s,a) &= Q_h^\star(s,a) - r_h(s,a) - [P_hV_{h+1}^k](s,a)\\ &= [P_h(V_{h+1}^\star - V_{h+1}^k)](s,a) \le 0.\end{aligned}$$

Combining the two branches, for every (s,a),

$$e_h^k(s,a) \le I_h^k(s,a)c_h^k(s,a). \tag{46}$$

Finally, use Lemma B.6 at the initial state, then (45) and (46):

$$\mathrm{Regret}(K) = \sum_{k=1}^K \big(V_1^\star(s_1) - V_1^{\pi^k}(s_1)\big)$$

$$\leq \sum_{k=1}^{K} \big(V_1^k(s_1) - V_1^{\pi^k}(s_1)\big)$$
$$\leq \sum_{k=1}^{K} \mathbb{E}\left[\sum_{h=1}^{H} I_h^k(s_h^k, a_h^k) c_h^k(s_h^k, a_h^k) \,\middle|\, \mathcal{H}_{k,0}\right].$$

The conditional expectations are with respect to the ordinary episode law; we do not condition that law on $\mathcal{E}$. The use of $\mathcal{E}$ is only to establish the pointwise comparisons for the already planned functions. Each trajectory cost lies in $[0, H^2]$, as claimed. □

Proof of Lemma B.10. We first establish a pointwise width recursion, then telescope its conditional expectations, and finally control predictably selected sums. The argument uses regression confidence and interval projection, but not exact optimism relative to $Q^\star$.

Let $\Delta V_h^k := V_h^k - \check{V}_h^k$. The projection and the ordered interval in Algorithm 1 imply $\Delta V_h^k \geq 0$ on every history. On $\mathcal{E}$, (35), which follows from Lemma B.3, yields

$$\begin{aligned} w_h^k(s,a) &= \hat{Q}_h^k(s,a) - \check{Q}_h^k(s,a) \\ &\leq \big(r_h(s,a) + [P_h V_{h+1}^k](s,a) + 2b_h^k(s,a)\big) \\ &\quad - \big(r_h(s,a) + [P_h \check{V}_{h+1}^k](s,a) - 2b_h^k(s,a)\big) \\ &= 4b_h^k(s,a) + [P_h \Delta V_{h+1}^k](s,a). \end{aligned} \tag{47}$$

If $4b_h^k(s,a) \leq H$, the bonus in this inequality equals $c_h^k(s,a)$. If $4b_h^k(s,a) > H$, use $w_h^k(s,a) \leq H = c_h^k(s,a)$ and $P_h \Delta V_{h+1}^k \geq 0$. Thus both cases give

$$w_h^k(s,a) \leq c_h^k(s,a) + [P_h \Delta V_{h+1}^k](s,a). \tag{48}$$

For the greedy action $a = \pi_h^k(s)$,

$$\begin{aligned} \Delta V_h^k(s) &= Q_h^k(s,a) - \max_{a'} \check{Q}_h^k(s,a') \\ &\leq Q_h^k(s,a) - \check{Q}_h^k(s,a) \leq \hat{Q}_h^k(s,a) - \check{Q}_h^k(s,a) = w_h^k(s,a). \end{aligned} \tag{49}$$

The first inequality evaluates the lower maximum at a; the second uses $Q_h^k \leq \hat{Q}_h^k$.

For $h < H$, applying (48) at the visited pair and then (49) at the next state gives

$$\begin{aligned} w_h^k(s_h^k, a_h^k) &\leq c_h^k(s_h^k, a_h^k) + \mathbb{E}[\Delta V_{h+1}^k(s_{h+1}^k) \mid \mathcal{H}_{k,h}] \\ &\leq c_h^k(s_h^k, a_h^k) + \mathbb{E}[w_{h+1}^k(s_{h+1}^k, a_{h+1}^k) \mid \mathcal{H}_{k,h}] \\ &\leq \mathbb{E}\big[c_h^k(s_h^k, a_h^k) + c_{h+1}^k(s_{h+1}^k, a_{h+1}^k) + \Delta V_{h+2}^k(s_{h+2}^k) \,\big|\, \mathcal{H}_{k,h}\big]. \end{aligned}$$

The last line uses the same recursion at stage $h+1$ and the tower property. Continuing through stage H and using $\Delta V_{H+1}^k = 0$ yields

$$w_h^k(s_h^k, a_h^k) \leq \mathbb{E}\left[\sum_{j=h}^{H} c_j^k(s_j^k, a_j^k) \,\middle|\, \mathcal{H}_{k,h}\right]. \tag{50}$$

In particular, at $h = H$ this says $w_H^k \leq c_H^k$; terminal-stage estimation uncertainty is included rather than set to zero.

Fix a stage h and a selection rule $J_k := \mathbb{1}\{k \in M\}$ that is $\mathcal{H}_{k,h}$-measurable. Define

$$X_k^{(h)} := J_k \sum_{j=h}^{H} c_j^k(s_j^k, a_j^k), \qquad m_k^{(h)} := \mathbb{E}[X_k^{(h)} \mid \mathcal{H}_{k,h}].$$

These selected costs satisfy $0 \leq X_k^{(h)} \leq H(H-h+1) \leq H^2$. Because J_k is known before the suffix is collected,

$$J_k w_h^k(s_h^k, a_h^k) \leq J_k \mathbb{E}\left[\sum_{j=h}^{H} c_j^k(s_j^k, a_j^k) \,\middle|\, \mathcal{H}_{k,h}\right]$$

$$= m_k^{(h)}.$$

For clarity, one may take the pre-observation filtration for cost k to be $\mathcal{H}_{k,h}$ and the next filtration to include the complete suffix and the next episode's prefix. These histories are nested, and $X_k^{(h)}$ is measurable by the next filtration. Thus the first inequality of Lemma E.3, with $B = H^2$, applies and gives

$$\begin{aligned}\sum_{k\in M} w_h^k(s_h^k, a_h^k) &\le \sum_{k=1}^{K} m_k^{(h)} \\ &\le 2\sum_{k=1}^{K} X_k^{(h)} + 2H^2\log(1/\eta) \\ &= 2\sum_{k\in M}\sum_{j=h}^{H} c_j^k(s_j^k, a_j^k) + 2H^2\log(1/\eta).\end{aligned}$$

The concentration event is defined for the costs without conditioning on $\mathcal{E}$; its intersection with $\mathcal{E}$ gives the stated result. This is a guarantee for the chosen predictable selection rule, not a simultaneous guarantee over arbitrary subsets selected after observing their suffixes. □

Proof of Lemma B.11. Fix h and let $n := |\mathcal{M}_h|$. If $n = 0$, the conclusion is immediate. Otherwise, the defining width threshold gives

$$\frac{\beta}{4} n \le \sum_{k\in\mathcal{M}_h} w_h^k(s_h^k, a_h^k). \tag{51}$$

Membership in $\mathcal{M}_h$ is known from the planning functions and (s_h^k, a_h^k) before the stage-h transition. We may therefore apply Lemma B.10 with this selection rule and failure probability δ/H. A union bound over h costs at most δ. On the resulting event intersected with $\mathcal{E}$,

$$\begin{aligned}\frac{\beta}{4} n &\overset{(a)}{\le} 2\sum_{j=h}^{H}\sum_{k\in\mathcal{M}_h} c_j^k(s_j^k, a_j^k) + 2H^2\ell \\ &\overset{(b)}{\le} 8\sqrt{C}\,\alpha\sqrt{n}\sum_{j=h}^{H}\sqrt{d_j} + 16\alpha\kappa(H-h+1)n + 2H^2\ell \\ &\overset{(c)}{\le} 8\sqrt{C}\,\alpha\sqrt{n(H-h+1)\sum_{j=h}^{H} d_j} + 16\alpha\kappa(H-h+1)n + 2H^2\ell \\ &\overset{(d)}{\le} 8\sqrt{C}\,\alpha H\sqrt{n d_{\mathcal{F}}} + 16\alpha + 2H^2\ell.\end{aligned} \tag{52}$$

Here (a) combines (51) with Lemma B.10 and $\log(H/\delta) \le \ell$. For (b), apply Lemma B.8 separately at each suffix stage j, always to the same realized subset $\mathcal{M}_h$; the lemma allows this subset to be data-dependent. Step (c) is Cauchy–Schwarz over the suffix stages. Finally, (d) uses

$$H - h + 1 \le H, \quad \sum_{j=h}^{H} d_j \le H d_{\mathcal{F}}, \quad n \le K, \quad \kappa HK = 1.$$

To solve (52) explicitly, set $A = 8\sqrt{C}\,\alpha H\sqrt{d_{\mathcal{F}}}$ and $B_0 = 16\alpha + 2H^2\ell$. Young's inequality gives

$$A\sqrt{n} \le \frac{\beta}{8} n + \frac{2A^2}{\beta}.$$

Substituting into $(\beta/4)n \le A\sqrt{n} + B_0$ and rearranging yields

$$\frac{\beta}{8} n \le \frac{2A^2}{\beta} + B_0,$$

$$n \le \frac{16A^2}{\beta^2} + \frac{8B_0}{\beta}$$
$$= \frac{1024C\alpha^2H^2d_{\mathcal{F}}}{\beta^2} + \frac{128\alpha}{\beta} + \frac{16H^2\ell}{\beta}. \tag{53}$$

The latter two terms are lower order under the appendix's parameter conventions. More explicitly, $\alpha \ge 1$, $d_{\mathcal{F}} \ge 1$, $\beta \le H$, and $\alpha^2 \ge H^2\ell$ imply

$$\frac{\alpha}{\beta} \le \frac{\alpha^2H^2d_{\mathcal{F}}}{\beta^2}, \qquad \frac{H^2\ell}{\beta} \le \frac{\alpha^2}{\beta} \le \frac{\alpha^2H^2d_{\mathcal{F}}}{\beta^2}.$$

Absorb these terms into an absolute constant C_1, and combine with the trivial bound $n \le K$. This proves

$$|\mathcal{M}_h| \le \min\{K, C_1H^2\alpha^2d_{\mathcal{F}}/\beta^2\}$$

simultaneously for all stages. The argument retains the sum of suffix-stage complexities and includes $j = H$ throughout. □

Proof of Lemma B.12. Apply Lemma B.8 at each stage with $M = \mathcal{M}_h$. Since $|\mathcal{M}_h| \le K$ and $\kappa HK = 1$,

$$S_M := \sum_{h=1}^{H} \sum_{k\in\mathcal{M}_h} c_h^k(s_h^k, a_h^k)$$
$$\le 4\sqrt{C}\,\alpha \sum_{h=1}^{H} \sqrt{|\mathcal{M}_h|d_h} + 8\alpha\kappa \sum_{h=1}^{H} |\mathcal{M}_h|$$
$$\le 4\sqrt{C}\,\alpha \sum_{h=1}^{H} \sqrt{|\mathcal{M}_h|d_h} + 8\alpha. \tag{54}$$

Use first the nontrivial cardinality estimate in Lemma B.11:

$$\sum_{h=1}^{H} \sqrt{|\mathcal{M}_h|d_h} \le \frac{\sqrt{C_1}H\alpha\sqrt{d_{\mathcal{F}}}}{\beta} \sum_{h=1}^{H} \sqrt{d_h}$$
$$\le \frac{\sqrt{C_1}H^2\alpha d_{\mathcal{F}}}{\beta}.$$

The final inequality is Cauchy–Schwarz, $\sum_h \sqrt{d_h} \le H\sqrt{d_{\mathcal{F}}}$. Substituting into (54) gives

$$S_M \le 4\sqrt{CC_1}\,\frac{H^2\alpha^2d_{\mathcal{F}}}{\beta} + 8\alpha.$$

Independently, use the trivial cardinality bound in the same expression:

$$S_M \le 4\sqrt{C}\,\alpha\sqrt{K} \sum_{h=1}^{H} \sqrt{d_h} + 8\alpha$$
$$\le 4\sqrt{C}\,\alpha H\sqrt{Kd_{\mathcal{F}}} + 8\alpha.$$

Both inequalities hold on the event in Lemma B.11, so we may take their minimum. This yields

$$S_M \lesssim \min\left\{\frac{H^2\alpha^2d_{\mathcal{F}}}{\beta}, \quad \alpha H\sqrt{Kd_{\mathcal{F}}}\right\} + \alpha,$$

as required. □

Proof of Lemma B.13. We first identify which reference errors can generate a narrow rejected interval, and then apply concentration to the true error indicators.

Suppose $\mathcal{E}$ holds and fix a visited pair with $Q_{\mathrm{ref};h}(s_h^k, a_h^k) = Q_h^\star(s_h^k, a_h^k)$ and $w_h^k(s_h^k, a_h^k) \le \beta/4$. For readability, suppress this pair's arguments below. Lemma B.4 gives $\check{Q}_h^k \le Q_h^\star \le \hat{Q}_h^k$. Therefore

$$\check{Q}_h^k = \hat{Q}_h^k - w_h^k \ge Q_h^\star - \beta/4 = Q_{\mathrm{ref};h} - \beta/4 \ge Q_{\mathrm{ref};h} - \beta/2,$$

$$\hat{Q}_h^k = \check{Q}_h^k + w_h^k \le Q_h^\star + \beta/4 = Q_{\mathrm{ref};h} + \beta/4 \le Q_{\mathrm{ref};h} + \beta/2.$$

Thus the entire interval is inside the reference neighborhood, so the test accepts. Taking the contrapositive and using the definition of $\mathcal{V}_h$ gives, on $\mathcal{E}$,

$$\mathbb{1}\{k \in \mathcal{V}_h\} \le Y_{k,h} := \mathbb{1}\{Q_{\mathrm{ref};h}(s_h^k, a_h^k) \neq Q_h^\star(s_h^k, a_h^k)\}. \tag{55}$$

Conditional on $\mathcal{H}_{k,0}$, the algorithm's policy π^k is fixed, and its stage-h pair has occupancy $d_h^{\pi^k}$. Consequently,

$$\begin{aligned}\mathbb{E}[Y_{k,h} \mid \mathcal{H}_{k,0}] &= \mathbb{E}_{(s,a)\sim d_h^{\pi^k}}\left[\mathbb{1}\{Q_{\mathrm{ref};h}(s,a) \neq Q_h^\star(s,a)\}\right] \\ &\le \rho. \end{aligned} \tag{56}$$

The last inequality is precisely Assumption 5.3, which holds for every policy and hence for the policy selected after any training history. Since $Y_{k,h} \in [0,1]$, the second inequality of Lemma E.3, with $B = 1$ and failure probability δ/H, gives

$$\begin{aligned}\sum_{k=1}^K Y_{k,h} &\le 2\sum_{k=1}^K \mathbb{E}[Y_{k,h} \mid \mathcal{H}_{k,0}] + \log(H/\delta) \\ &\le 2\rho K + \log(H/\delta).\end{aligned}$$

A union bound makes this simultaneous for all h with failure probability at most δ. Intersect this event with $\mathcal{E}$ and apply (55) to conclude

$$|\mathcal{V}_h| \le \sum_{k=1}^K Y_{k,h} \le 2\rho K + \log(H/\delta).$$

Concentration was applied to $Y_{k,h}$ without conditioning on the future confidence event. If $\rho = 0$, the nonnegative variable $Y_{k,h}$ has zero conditional expectation by (56), so it is zero almost surely. A finite union over h, k then gives $|\mathcal{V}_h| = 0$ on $\mathcal{E}$. □

Proof of Lemma B.14. By Lemma B.8, followed by Lemma B.13,

$$\begin{aligned}S_V &:= \sum_{h=1}^H \sum_{k\in\mathcal{V}_h} c_h^k(s_h^k, a_h^k) \\ &\le 4\sqrt{C}\,\alpha \sum_{h=1}^H \sqrt{|\mathcal{V}_h| d_h} + 8\alpha\kappa \sum_{h=1}^H |\mathcal{V}_h| \\ &\le 4\sqrt{C}\,\alpha \sqrt{2\rho K + \log(H/\delta)} \sum_{h=1}^H \sqrt{d_h} + 8\alpha\kappa HK \\ &\le 4\sqrt{C}\,\alpha H \sqrt{d_{\mathcal{F}}(2\rho K + \log(H/\delta))} + 8\alpha \\ &\lesssim \alpha H \sqrt{d_{\mathcal{F}}(\rho K + \ell)} + \alpha. \end{aligned} \tag{57}$$

The third line also uses $|\mathcal{V}_h| \le K$ for the rounding term; the fourth uses Cauchy–Schwarz and $\kappa HK = 1$; and the last uses $\log(H/\delta) \le \ell$.

For the additional assertion, if a pair passes the test, Lemma B.5 gives $Q_h^k = Q_h^\star$ and Lemma B.6 gives $V_{h+1}^k \ge V_{h+1}^\star$. Hence

$$Q_h^k(s,a) - r_h(s,a) - [P_h V_{h+1}^k](s,a) = [P_h(V_{h+1}^\star - V_{h+1}^k)](s,a) \le 0.$$

Its positive part is therefore zero. This remains true whether or not the accepted interval belongs to the wide set $\mathcal{M}_h$. □

C Proof of Theorem A.2

C.1 Sets

The algorithm now receives $\tilde{Q}_{\mathrm{ref}}$, and its acceptance test is centered at this observed reference. Define

$$\widetilde{\mathcal{M}}_h := \{k : w_h^k(s_h^k, a_h^k) > \beta/4\}, \tag{58}$$

$$\widetilde{\mathcal{V}}_h := \{k \notin \widetilde{\mathcal{M}}_h : [\check{Q}_h^k(s_h^k, a_h^k), \hat{Q}_h^k(s_h^k, a_h^k)] \not\subseteq [\tilde{Q}_{\mathrm{ref};h}(s_h^k, a_h^k) - \beta/2, \tilde{Q}_{\mathrm{ref};h}(s_h^k, a_h^k) + \beta/2]\}. \tag{59}$$

These sets use the actual algorithmic test, not an enlarged neighborhood around the unobserved ideal reference. Let $\widetilde{I}_h^k$ indicate failure of this test. The regression event $\mathcal{E}$ and the nonnegative costs c_h^k are defined as in Appendix B, conditional on the fixed observed reference. The finite-family concentration proof is unchanged.

C.2 Optimism Lemmas under Misspecification

Lemma C.1 (Certification under small misspecification)**.** *On $\mathcal{E}$, if $0 \le \tau \le \beta/8$, every accepted pair satisfies $Q_{\mathrm{ref};h}(s,a) = Q_h^\star(s,a)$ and*

$$|Q_h^k(s,a) - Q_h^\star(s,a)| \le \tau.$$

Here Q_h^k is the projection of $\tilde{Q}_{\mathrm{ref};h}$ onto the confidence interval, as specified in Algorithm 1.

Lemma C.2 (Approximate optimism)**.** *On $\mathcal{E}$, pessimism $\check{Q}_h^k \le Q_h^\star$ holds for every fixed observed reference. If $\tau \le \beta/8$, then*

$$\hat{Q}_h^k \ge Q_h^\star - \tau, \qquad Q_h^k \ge Q_h^\star - \tau, \qquad V_h^k \ge V_h^\star - \tau.$$

For an arbitrary observed reference, the weaker bounds

$$Q_h^k \ge Q_h^\star - (H - h + 1)\beta/2, \qquad V_h^k \ge V_h^\star - (H - h + 1)\beta/2$$

hold. These statements do not require the learner to know τ.

C.3 Misspecification Lemmas

Lemma C.3. *If $\tau \le \beta/8$, then on $\mathcal{E}$ and an additional event of probability at least $1 - \delta$,*

$$|\widetilde{\mathcal{V}}_h| \le 2\rho K + \log(H/\delta), \qquad h \in [H].$$

Lemma C.4 (Decomposition)**.** *On $\mathcal{E}$, for $\tau \le \beta/8$,*

$$\mathrm{Regret}(K) \le (2H+1)K\tau + \sum_{k=1}^{K} \mathbb{E}\left[\sum_{h=1}^{H} \widetilde{I}_h^k(s_h^k, a_h^k) c_h^k(s_h^k, a_h^k) \,\middle|\, \mathcal{H}_{k,0}\right].$$

For arbitrary $\tilde{Q}_{\mathrm{ref}}$, we also have both

$$\mathrm{Regret}(K) \le \frac{3}{4}HK\beta + \sum_k \mathbb{E}\left[\sum_h \mathbb{1}\{k \in \widetilde{\mathcal{M}}_h\} c_h^k \,\middle|\, \mathcal{H}_{k,0}\right],$$

$$\mathrm{Regret}(K) \le \frac{1}{2}HK\beta + \sum_k \mathbb{E}\left[\sum_h c_h^k \,\middle|\, \mathcal{H}_{k,0}\right],$$

where trajectory costs are evaluated at the visited state-action pairs.

Lemma C.5 (Upper bound on $|\widetilde{\mathcal{M}}_h|$)**.** *For any observed reference, on $\mathcal{E}$ and an additional event of probability at least $1 - \delta$,*

$$|\widetilde{\mathcal{M}}_h| \le \min\{K, C_1 H^2 \alpha^2 d_{\mathcal{F}}/\beta^2\}, \qquad h \in [H]. \tag{60}$$

Lemma C.6. *On the events of Lemma C.5,*

$$\sum_h \sum_{k \in \widetilde{\mathcal{M}}_h} c_h^k(s_h^k, a_h^k) \lesssim \min\{H^2\alpha^2 d_{\mathcal{F}}/\beta, \alpha H\sqrt{K d_{\mathcal{F}}}\} + \alpha.$$

Lemma C.7. *If $\tau \le \beta/8$, then on the events of Lemma C.3,*

$$\sum_h \sum_{k \in \widetilde{\mathcal{V}}_h} c_h^k(s_h^k, a_h^k) \lesssim \alpha H\sqrt{d_{\mathcal{F}}(\rho K + \ell)} + \alpha.$$

For any observed reference, the Bellman residual on a narrow interval is at most $\beta/4$, regardless of whether the reference is accepted.

C.4 Detailed Proof

Proof of Theorem A.2. The proof distinguishes two ranges of τ. When the error is small, certification can still identify exact entries of the ideal reference. For the remaining range, a bound proportional to $HK\beta$ suffices because $\beta < 8\tau$. Throughout, the algorithm receives only $\widetilde{Q}_{\mathrm{ref}}$; this case distinction is for the analysis and is not used by the learner.

First, consider $0 < \tau \le \beta/8$. Define the nonnegative episode cost

$$\widetilde{X}_k := \sum_{h=1}^H \widetilde{I}_h^k(s_h^k, a_h^k) c_h^k(s_h^k, a_h^k), \qquad \widetilde{m}_k := \mathbb{E}[\widetilde{X}_k \mid \mathcal{H}_{k,0}].$$

As before, $0 \le \widetilde{X}_k \le H^2$ on every history. Apply Lemma E.3 with $B = H^2$ and failure probability δ. Intersect this event with $\mathcal{E}$ and the additional events of Lemmas C.5 and C.3. The union bound gives total failure probability at most $2\delta + \delta + \delta + \delta = 5\delta$. On this intersection,

$$\begin{aligned}
\mathrm{Regret}(K) &\overset{(a)}{\le} (2H+1)K\tau + \sum_{k=1}^K \widetilde{m}_k \\
&\overset{(b)}{\le} (2H+1)K\tau + 2\sum_{k=1}^K \widetilde{X}_k + 2H^2\ell \\
&\overset{(c)}{\le} (2H+1)K\tau + 2\sum_{h=1}^H \sum_{k \in \widetilde{\mathcal{M}}_h} c_h^k(s_h^k, a_h^k) \\
&\qquad + 2\sum_{h=1}^H \sum_{k \in \widetilde{\mathcal{V}}_h} c_h^k(s_h^k, a_h^k) + 2H^2\ell \\
&\overset{(d)}{\lesssim} HK\tau + \frac{H^2\alpha^2 d_{\mathcal{F}}}{\beta} + \alpha H\sqrt{d_{\mathcal{F}}(\rho K + \ell)} + \alpha + H^2\ell. \qquad (61)
\end{aligned}$$

Here (a) uses the first decomposition in Lemma C.4; (b) uses the cost concentration inequality; (c) uses (58)–(59); and (d) uses Lemmas C.6 and C.7, together with $(2H+1)K\tau \le 3HK\tau$. Applying $\sqrt{\rho K + \ell} \le \sqrt{\rho K} + \sqrt{\ell}$ and the three explicit remainder bounds in (25) gives

$$\mathrm{Regret}(K) \lesssim \frac{H^2\alpha^2 d_{\mathcal{F}}}{\beta} + \alpha H\sqrt{\rho K d_{\mathcal{F}}} + HK\tau. \qquad (62)$$

For the online branch, use $\widetilde{I}_h^k \le 1$ directly in (b), then apply Lemma B.7:

$$\begin{aligned}
\mathrm{Regret}(K) &\le (2H+1)K\tau + 2\sum_{h=1}^H \sum_{k=1}^K c_h^k(s_h^k, a_h^k) + 2H^2\ell \\
&\le (2H+1)K\tau + 8\sqrt{C}\,\alpha\sqrt{K}\sum_{h=1}^H \sqrt{d_h} + 16\alpha\kappa KH + 2H^2\ell
\end{aligned}$$

$$\leq 3HK\tau + 8\sqrt{C}\,\alpha H\sqrt{Kd_{\mathcal{F}}} + 16\alpha + 2H^2\ell. \tag{63}$$

For $K \geq \ell$, the last two terms are absorbed exactly as in (27). For $K < \ell$, the deterministic bound (28) applies to the true regret, whether or not the planning values are optimistic. Thus

$$\mathrm{Regret}(K) \lesssim \alpha H\sqrt{Kd_{\mathcal{F}}} + HK\tau. \tag{64}$$

Both (62) and (64) hold on the same event; their minimum proves the required bound for this error range.

Next, consider $\beta/8 < \tau \leq \beta/2$. In this range, we do not rely on certification of the ideal reference or on an inaccurate-region count. Define the wide-interval and all-bonus costs

$$X_k^M := \sum_{h=1}^{H} \mathbb{1}\{k \in \widetilde{\mathcal{M}}_h\} c_h^k(s_h^k, a_h^k),$$
$$X_k^{\mathrm{all}} := \sum_{h=1}^{H} c_h^k(s_h^k, a_h^k).$$

Both costs lie in $[0, H^2]$. Apply the first inequality of Lemma E.3 to each cost sequence with failure probability $\delta/2$. Together with $\mathcal{E}$ and the additional wide-interval event in Lemma C.5, the total failure probability is at most $2\delta + \delta + \delta/2 + \delta/2 = 4\delta \leq 5\delta$.

The first generic decomposition of Lemma C.4 and Lemma C.6 imply

$$\begin{aligned}
\mathrm{Regret}(K) &\leq \frac{3}{4}HK\beta + \sum_{k=1}^{K} \mathbb{E}[X_k^M \mid \mathcal{H}_{k,0}] \\
&\leq \frac{3}{4}HK\beta + 2\sum_{k=1}^{K} X_k^M + 2H^2\log(2/\delta) \\
&\lesssim HK\beta + \frac{H^2\alpha^2 d_{\mathcal{F}}}{\beta} + \alpha + H^2\ell \\
&\lesssim HK\beta + \frac{H^2\alpha^2 d_{\mathcal{F}}}{\beta}.
\end{aligned} \tag{65}$$

The last line uses (25). Independently, the second generic decomposition in Lemma C.4 and Lemma B.7 give

$$\begin{aligned}
\mathrm{Regret}(K) &\leq \frac{1}{2}HK\beta + \sum_{k=1}^{K} \mathbb{E}[X_k^{\mathrm{all}} \mid \mathcal{H}_{k,0}] \\
&\leq \frac{1}{2}HK\beta + 2\sum_{k=1}^{K} X_k^{\mathrm{all}} + 2H^2\log(2/\delta) \\
&\lesssim HK\beta + \alpha H\sqrt{Kd_{\mathcal{F}}} + \alpha + H^2\ell.
\end{aligned} \tag{66}$$

For $K \geq \ell$ absorb the last two terms into $\alpha H\sqrt{Kd_{\mathcal{F}}}$, and for $K < \ell$ use (28). Combining the two resulting bounds gives

$$\begin{aligned}
\mathrm{Regret}(K) &\lesssim \min\left\{\frac{H^2\alpha^2 d_{\mathcal{F}}}{\beta},\quad \alpha H\sqrt{Kd_{\mathcal{F}}}\right\} + HK\beta \\
&\lesssim \min\left\{\frac{H^2\alpha^2 d_{\mathcal{F}}}{\beta} + \alpha H\sqrt{\rho K d_{\mathcal{F}}},\quad \alpha H\sqrt{Kd_{\mathcal{F}}}\right\} + HK\tau.
\end{aligned} \tag{67}$$

For the second inequality, the added square-root term is nonnegative, and $\beta < 8\tau$ in this range. The absolute factor eight is absorbed in the implicit constant.

Finally, the two cases cover the full interval $0 < \tau \leq \beta/2$. Substitute $\alpha = c_\alpha H\sqrt{\mathcal{L}_K}$ as in the proof of Theorem 6.2. This yields the same two leading regret terms, with the additional $\mathcal{O}(HK\tau)$ term, and proves the result with probability at least $1 - 5\delta$ under the appendix's complexity convention. Neither case requires the algorithm to know τ. □

C.5 Detailed Proof for Supporting Lemmas

Joint proof of Lemmas C.1 and C.2. Fix an episode and work on $\mathcal{E}$. We first prove pessimism without any restriction on the observed reference. We then establish small-error certification and approximate optimism jointly, followed by the weaker bound for arbitrary references.

At stage $H+1$, $\check{V}^k_{H+1} = V^\star_{H+1} = 0$. Suppose $\check{V}^k_{h+1} \le V^\star_{h+1}$. By Lemma B.3 and the Bellman optimality equation,

$$\begin{aligned}\check{f}^k_h(s,a) - b^k_h(s,a) &\le r_h(s,a) + [P_h \check{V}^k_{h+1}](s,a) \\ &\le r_h(s,a) + [P_h V^\star_{h+1}](s,a) = Q^\star_h(s,a).\end{aligned}$$

As $Q^\star_h(s,a) \in [0, H_h]$, clipping the left side to $[0, H_h]$ preserves its upper bound. Therefore $\check{Q}^k_h \le Q^\star_h$ and, after maximizing over actions, $\check{V}^k_h \le V^\star_h$. Backward induction proves these inequalities for every fixed observed reference.

Assume $0 \le \tau \le \beta/8$. The terminal continuation is zero, so $V^k_{H+1} \ge V^\star_{H+1} - \tau$. Suppose inductively that

$$V^k_{h+1}(s') \ge V^\star_{h+1}(s') - \tau \qquad \text{for every } s'. \tag{68}$$

By (35) and the fact that each $P_h(\cdot \mid s,a)$ is a probability distribution,

$$\begin{aligned}\hat{Q}^k_h(s,a) &\ge r_h(s,a) + [P_h V^k_{h+1}](s,a) \\ &\ge r_h(s,a) + [P_h V^\star_{h+1}](s,a) - \tau = Q^\star_h(s,a) - \tau.\end{aligned} \tag{69}$$

Notice that this is approximate, rather than exact, optimism.

If the test accepts at (s,a), its containment condition gives

$$\tilde{Q}_{\mathrm{ref};h} - \beta/2 \le \check{Q}^k_h \le \hat{Q}^k_h \le \tilde{Q}_{\mathrm{ref};h} + \beta/2,$$

where the common arguments (s,a) are suppressed. Using the pessimism established above and (69), respectively, we obtain both sides of the following interval:

$$\begin{aligned}\tilde{Q}_{\mathrm{ref};h} &\le \check{Q}^k_h + \beta/2 \le Q^\star_h + \beta/2, \\ \tilde{Q}_{\mathrm{ref};h} &\ge \hat{Q}^k_h - \beta/2 \ge Q^\star_h - \tau - \beta/2.\end{aligned} \tag{70}$$

Definition A.1 states $|Q_{\mathrm{ref};h} - \tilde{Q}_{\mathrm{ref};h}| \le \tau$. Combining this with (70) yields

$$\begin{aligned}Q^\star_h - \beta/2 - 2\tau \le Q_{\mathrm{ref};h} \le Q^\star_h + \beta/2 + \tau, \\ |Q_{\mathrm{ref};h} - Q^\star_h| \le \beta/2 + 2\tau \le 3\beta/4 < \beta.\end{aligned}$$

By Definition 5.1, the ideal reference must therefore be exact: $Q_{\mathrm{ref};h} = Q^\star_h$. It follows that $|\tilde{Q}_{\mathrm{ref};h} - Q^\star_h| \le \tau$.

It remains to check that projection preserves this error bound even when $Q^\star_h$ is not inside the interval. Let $l = \check{Q}^k_h$, $u = \hat{Q}^k_h$, and $q = \tilde{Q}_{\mathrm{ref};h}$. The inequalities established above say

$$l \le Q^\star_h, \qquad u \ge Q^\star_h - \tau, \qquad Q^\star_h - \tau \le q \le Q^\star_h + \tau.$$

The two endpoints needed for projection satisfy

$$\begin{aligned}&\max\{l,q\} \ge q \ge Q^\star_h - \tau, \qquad u \ge Q^\star_h - \tau, \\ &\max\{l,q\} \le Q^\star_h + \tau.\end{aligned}$$

Therefore

$$Q^\star_h - \tau \le \min\{u, \max\{l,q\}\} = \Pi_{[l,u]}(q) = Q^k_h \le Q^\star_h + \tau. \tag{71}$$

If the test rejects, the algorithm instead sets $Q^k_h = \hat{Q}^k_h \ge Q^\star_h - \tau$ by (69). Thus both branches satisfy $Q^k_h(s,a) \ge Q^\star_h(s,a) - \tau$. Maximizing over actions gives

$$V^k_h(s) = \max_a Q^k_h(s,a) \ge \max_a \big(Q^\star_h(s,a) - \tau\big) = V^\star_h(s) - \tau.$$

This closes the induction and proves small-error certification and all small-error claims in Lemma C.2.

For a possibly arbitrary $\tilde{Q}_{\mathrm{ref}}$, we prove by backward induction that $V_h^k \geq V_h^\star - H_h\beta/2$. The terminal case again holds. Assume $V_{h+1}^k \geq V_{h+1}^\star - (H-h)\beta/2$. Equation (35) gives

$$\hat{Q}_h^k \geq r_h + P_h V_{h+1}^k \geq Q_h^\star - (H-h)\beta/2.$$

On an accepted pair, the actual inclusion test implies $\tilde{Q}_{\mathrm{ref};h} \geq \hat{Q}_h^k - \beta/2$. Thus

$$\begin{aligned} Q_h^k &= \min\{\hat{Q}_h^k, \max\{\check{Q}_h^k, \tilde{Q}_{\mathrm{ref};h}\}\} \\ &\geq \min\{\hat{Q}_h^k, \hat{Q}_h^k - \beta/2\} = \hat{Q}_h^k - \beta/2 \\ &\geq Q_h^\star - (H-h+1)\beta/2. \end{aligned}$$

On a rejected pair, $Q_h^k = \hat{Q}_h^k$ satisfies an even stronger bound. Maximizing over actions proves the value-function inequality and completes the induction. This last step does not use τ, exact certification, or the coverage assumption. □

Proof of Lemma C.3. Work first on $\mathcal{E}$ and suppose $\tau \leq \beta/8$. Consider a visited pair at which the ideal reference is exact, $Q_{\mathrm{ref};h} = Q_h^\star$, and the interval is narrow, $w_h^k \leq \beta/4$. We show separately that both endpoints satisfy the actual acceptance test centered at $\tilde{Q}_{\mathrm{ref};h}$. Suppress the visited pair's arguments. For the lower endpoint,

$$\begin{aligned} \check{Q}_h^k &= \hat{Q}_h^k - w_h^k \\ &\overset{(a)}{\geq} \hat{Q}_h^k - \beta/4 \\ &\overset{(b)}{\geq} Q_h^\star - \tau - \beta/4 \\ &\overset{(c)}{\geq} \tilde{Q}_{\mathrm{ref};h} - 2\tau - \beta/4 \\ &\overset{(d)}{\geq} \tilde{Q}_{\mathrm{ref};h} - \beta/2. \end{aligned} \tag{72}$$

Here (a) is the narrow-width condition; (b) is Lemma C.2; (c) follows from $Q_{\mathrm{ref};h} = Q_h^\star$ and Definition A.1; and (d) uses $2\tau \leq \beta/4$. For the upper endpoint,

$$\begin{aligned} \hat{Q}_h^k &= \check{Q}_h^k + w_h^k \\ &\overset{(a)}{\leq} \check{Q}_h^k + \beta/4 \\ &\overset{(b)}{\leq} Q_h^\star + \beta/4 \\ &\overset{(c)}{\leq} \tilde{Q}_{\mathrm{ref};h} + \tau + \beta/4 \\ &\overset{(d)}{\leq} \tilde{Q}_{\mathrm{ref};h} + \beta/2. \end{aligned} \tag{73}$$

The justifications are the width condition, pessimism in Lemma C.2, Definition A.1, and $\tau \leq \beta/8 \leq \beta/4$, respectively. Thus the actual test accepts whenever the ideal reference is exact and the interval is narrow. Consequently, on $\mathcal{E}$,

$$\mathbb{1}\{k \in \widetilde{\mathcal{V}}_h\} \leq Y_{k,h} := \mathbb{1}\{Q_{\mathrm{ref};h}(s_h^k, a_h^k) \neq Q_h^\star(s_h^k, a_h^k)\}. \tag{74}$$

We now apply concentration to $Y_{k,h}$, not to an indicator conditioned on $\mathcal{E}$. Even though π^k is selected using the observed reference, Assumption 5.3 applies to every policy. Therefore

$$\mathbb{E}[Y_{k,h} \mid \mathcal{H}_{k,0}] = \mathbb{E}_{(s,a)\sim d_h^{\pi^k}}[\mathbb{1}\{Q_{\mathrm{ref};h}(s,a) \neq Q_h^\star(s,a)\}] \leq \rho.$$

By the second inequality of Lemma E.3, with $B = 1$ and failure probability δ/H,

$$\begin{aligned} \sum_{k=1}^K Y_{k,h} &\leq 2\sum_{k=1}^K \mathbb{E}[Y_{k,h} \mid \mathcal{H}_{k,0}] + \log(H/\delta) \\ &\leq 2\rho K + \log(H/\delta). \end{aligned}$$

A union bound over stages costs at most δ. Intersecting this event with $\mathcal{E}$ and using (74) proves the lemma. This is the same true-error counting mechanism as in Lemma B.13, with the endpoint calculations above supplying the required misspecified-reference implication. □

Proof of Lemma C.4. Fix an episode k, and define $e_h^k := Q_h^k - r_h - P_h V_{h+1}^k$. We first derive residual bounds valid for any observed reference.

On $\mathcal{E}$, projection gives $Q_h^k \le \hat{Q}_h^k$, and (35) gives $\hat{Q}_h^k \le r_h + P_h V_{h+1}^k + 2b_h^k$. Thus

$$e_h^k \le 2b_h^k.$$

Since $Q_h^k \le H$ and $r_h + P_h V_{h+1}^k \ge 0$, we also have $e_h^k \le H$, yielding $e_h^k \le \min\{2b_h^k, H\} \le c_h^k$. For the width bound, (35) implies

$$\check{Q}_h^k \le r_h + P_h \check{V}_{h+1}^k \le r_h + P_h V_{h+1}^k,$$

where the second comparison uses $\check{V}_{h+1}^k \le V_{h+1}^k$ from projection. Consequently,

$$e_h^k = Q_h^k - (r_h + P_h V_{h+1}^k) \le Q_h^k - \check{Q}_h^k \le \hat{Q}_h^k - \check{Q}_h^k = w_h^k. \tag{75}$$

We have established, without a restriction on τ,

$$e_h^k \le c_h^k, \qquad e_h^k \le w_h^k. \tag{76}$$

When $\tau \le \beta/8$, certification gives a sharper residual bound. On a pair accepted by the reference test, Lemma C.1 gives $Q_h^k \le Q_h^\star + \tau$, and Lemma C.2 gives $V_{h+1}^k \ge V_{h+1}^\star - \tau$. Therefore

$$\begin{aligned} e_h^k(s,a) &\le Q_h^\star(s,a) + \tau - r_h(s,a) - [P_h V_{h+1}^k](s,a) \\ &\le Q_h^\star(s,a) + \tau - r_h(s,a) - [P_h V_{h+1}^\star](s,a) + \tau \\ &= 2\tau. \end{aligned} \tag{77}$$

The final equality uses the Bellman optimality equation. On a rejected pair, (76) gives $e_h^k \le c_h^k$. Combining the two branches produces the pointwise bound

$$e_h^k(s,a) \le \widetilde{I}_h^k(s,a) c_h^k(s,a) + 2\tau. \tag{78}$$

The conditional Bellman identity (45), proved in the proof of Lemma B.9, does not require optimism and hence still holds. Lemma C.2 further gives $V_1^\star(s_1) \le V_1^k(s_1) + \tau$ in the small-error regime. Thus

$$\begin{aligned} V_1^\star(s_1) - V_1^{\pi^k}(s_1) &\le \tau + V_1^k(s_1) - V_1^{\pi^k}(s_1) \\ &= \tau + \mathbb{E}\left[\sum_{h=1}^H e_h^k(s_h^k, a_h^k) \,\middle|\, \mathcal{H}_{k,0}\right] \\ &\le (2H+1)\tau + \mathbb{E}\left[\sum_{h=1}^H \widetilde{I}_h^k(s_h^k, a_h^k) c_h^k(s_h^k, a_h^k) \,\middle|\, \mathcal{H}_{k,0}\right]. \end{aligned} \tag{79}$$

Summing over k proves the first assertion.

For an arbitrary observed reference, the generic part of Lemma C.2 yields $V_1^\star(s_1) \le V_1^k(s_1) + H\beta/2$. If $w_h^k \le \beta/4$, (75) gives $e_h^k \le \beta/4$. If $w_h^k > \beta/4$, use $e_h^k \le c_h^k$ instead. These two cases imply

$$e_h^k(s,a) \le \mathbb{1}\{w_h^k(s,a) > \beta/4\} c_h^k(s,a) + \beta/4.$$

Along the realized trajectory the indicator equals $\mathbb{1}\{k \in \widetilde{\mathcal{M}}_h\}$, so conditional Bellman telescoping now gives

$$\begin{aligned} V_1^\star(s_1) - V_1^{\pi^k}(s_1) &\le \frac{H\beta}{2} + \mathbb{E}\left[\sum_{h=1}^H e_h^k(s_h^k, a_h^k) \,\middle|\, \mathcal{H}_{k,0}\right] \\ &\le \frac{H\beta}{2} + \frac{H\beta}{4} + \mathbb{E}\left[\sum_{h=1}^H \mathbb{1}\{k \in \widetilde{\mathcal{M}}_h\} c_h^k(s_h^k, a_h^k) \,\middle|\, \mathcal{H}_{k,0}\right] \\ &= \frac{3}{4} H\beta + \mathbb{E}\left[\sum_{h=1}^H \mathbb{1}\{k \in \widetilde{\mathcal{M}}_h\} c_h^k(s_h^k, a_h^k) \,\middle|\, \mathcal{H}_{k,0}\right]. \end{aligned}$$

Alternatively, using only $e_h^k \le c_h^k$ gives

$$V_1^\star(s_1) - V_1^{\pi^k}(s_1) \le \frac{1}{2}H\beta + \mathbb{E}\left[\sum_{h=1}^{H} c_h^k(s_h^k, a_h^k) \,\middle|\, \mathcal{H}_{k,0}\right].$$

Summing these two inequalities separately over episodes proves both remaining assertions. The nonnegative costs inside all decompositions are defined under the ordinary interaction law, without conditioning on the future event $\mathcal{E}$. □

Proof of Lemmas C.5 and C.6. We give the details to identify why the wide-interval analysis remains valid even without exact optimism. For any observed reference, Algorithm 1 still guarantees $\check{Q}_h^k \le Q_h^k \le \hat{Q}_h^k$ by projection. The confidence sandwich (35) therefore applies on $\mathcal{E}$, as do the width recursion and predictable-suffix bound of Lemma B.10.

Fix h and put $n_h := |\widetilde{\mathcal{M}}_h|$. Its membership indicator is $\mathcal{H}_{k,h}$-measurable. Apply Lemma B.10 with failure probability δ/H and then Lemma B.8 at every suffix stage. On an additional event with failure probability at most δ across all stages,

$$\begin{aligned}
\frac{\beta}{4}n_h &\le \sum_{k\in\widetilde{\mathcal{M}}_h} w_h^k(s_h^k, a_h^k) \\
&\le 2\sum_{j=h}^{H}\sum_{k\in\widetilde{\mathcal{M}}_h} c_j^k(s_j^k, a_j^k) + 2H^2\ell \\
&\le 8\sqrt{C}\,\alpha\sqrt{n_h}\sum_{j=h}^{H}\sqrt{d_j} + 16\alpha\kappa(H-h+1)n_h + 2H^2\ell \\
&\le 8\sqrt{C}\,\alpha H\sqrt{n_h d_{\mathcal{F}}} + 16\alpha + 2H^2\ell. \qquad (80)
\end{aligned}$$

The first inequality is the definition of $\widetilde{\mathcal{M}}_h$; the next two use the cited lemmas; and the last uses Cauchy–Schwarz, $n_h \le K$, and $\kappa HK = 1$. Set $A = 8\sqrt{C}\,\alpha H\sqrt{d_{\mathcal{F}}}$ and $B_0 = 16\alpha + 2H^2\ell$. As in (53), Young's inequality gives

$$\begin{aligned}
\frac{\beta}{4}n_h &\le A\sqrt{n_h} + B_0 \le \frac{\beta}{8}n_h + \frac{2A^2}{\beta} + B_0, \\
n_h &\le \frac{16A^2}{\beta^2} + \frac{8B_0}{\beta} \le \frac{C_1 H^2\alpha^2 d_{\mathcal{F}}}{\beta^2}.
\end{aligned}$$

The last step uses the same explicitly verified remainder absorption as in the proof of Lemma B.11. Together with $n_h \le K$, this proves Lemma C.5 with the same absolute constant chosen large enough.

For the cost assertion, apply Lemma B.8 again:

$$\begin{aligned}
\sum_{h=1}^{H}\sum_{k\in\widetilde{\mathcal{M}}_h} c_h^k(s_h^k, a_h^k) &\le 4\sqrt{C}\,\alpha\sum_{h=1}^{H}\sqrt{n_h d_h} + 8\alpha\kappa\sum_{h=1}^{H} n_h \\
&\le 4\sqrt{C}\,\alpha\min\left\{\sqrt{K}, \frac{\sqrt{C_1}H\alpha\sqrt{d_{\mathcal{F}}}}{\beta}\right\}\sum_{h=1}^{H}\sqrt{d_h} + 8\alpha \\
&\lesssim \min\left\{\alpha H\sqrt{K d_{\mathcal{F}}}, \frac{H^2\alpha^2 d_{\mathcal{F}}}{\beta}\right\} + \alpha.
\end{aligned}$$

The second line uses the two uniform bounds on n_h and $\kappa HK = 1$; the last uses $\sum_h \sqrt{d_h} \le H\sqrt{d_{\mathcal{F}}}$. This proves Lemma C.6. No step used the accuracy of $\tilde{Q}_{\mathrm{ref}}$, the value of τ, or the coverage coefficient ρ. □

Proof of Lemma C.7. For $\tau \le \beta/8$, Lemma B.8, followed by Lemma C.3, gives

$$\sum_{h=1}^{H}\sum_{k\in\widetilde{\mathcal{V}}_h} c_h^k(s_h^k, a_h^k)$$

$$
\begin{aligned}
&\le 4\sqrt{C}\,\alpha \sum_{h=1}^{H} \sqrt{|\widetilde{\mathcal{V}}_h| d_h} + 8\alpha\kappa \sum_{h=1}^{H} |\widetilde{\mathcal{V}}_h| \\
&\le 4\sqrt{C}\,\alpha \sqrt{2\rho K + \log(H/\delta)} \sum_{h=1}^{H} \sqrt{d_h} + 8\alpha\kappa HK \\
&\le 4\sqrt{C}\,\alpha H \sqrt{d_{\mathcal{F}}(2\rho K + \log(H/\delta))} + 8\alpha \\
&\lesssim \alpha H \sqrt{d_{\mathcal{F}}(\rho K + \ell)} + \alpha.
\end{aligned}
$$

The third displayed line also uses $|\widetilde{\mathcal{V}}_h| \le K$ for the rounding term. The next uses Cauchy–Schwarz and $\kappa HK = 1$; the last uses $\log(H/\delta) \le \ell$.

For the second assertion, no small-error condition is needed. Equation (75), established in the proof of Lemma C.4, gives

$$
\begin{aligned}
Q_h^k(s,a) - r_h(s,a) - [P_h V_{h+1}^k](s,a) &\le Q_h^k(s,a) - \check{Q}_h^k(s,a) \\
&\le \hat{Q}_h^k(s,a) - \check{Q}_h^k(s,a) = w_h^k(s,a).
\end{aligned}
$$

Whenever the interval is narrow, its final width is at most $\beta/4$. This proves the claimed residual bound regardless of which branch of the reference test is selected. □

D Proof of Theorem 4.1

D.1 Construction

We use the time-inhomogeneous two-state absorbing construction already described in the lower-bound argument, following the construction strategy in Zhou et al. (2021). Let $d \ge 3$, $m = d - 2$, and $\mathcal{A} = \{-1, +1\}^m$. The states are a transient state x and a good absorbing state x_G, and every episode starts at x. The rewards are

$$
r_h(x,a) = 0, \qquad r_h(x_G,a) = 1.
$$

For $0 < \epsilon \le \zeta H/384$, define

$$
p_0 = \frac{1}{2H}, \qquad \Delta = \frac{1}{mH^2}, \qquad \gamma = 128\epsilon, \qquad u = \gamma\Delta = \frac{128\epsilon}{mH^2}. \tag{81}
$$

For every stage, select a sign vector $\theta_h \in \{-1,+1\}^m$ and let $\boldsymbol{\mu}_h = \Delta\theta_h$. The transition probabilities from x are

$$
\begin{aligned}
P_h(x_G \mid x, a) &= p_0 + \gamma\langle \boldsymbol{\mu}_h, a\rangle = p_0 + u\langle \theta_h, a\rangle, \\
P_h(x \mid x, a) &= 1 - p_0 - u\langle \theta_h, a\rangle,
\end{aligned}
$$

while $P_h(x_G \mid x_G, a) = 1$.

We first check that these are valid probabilities. Since $|\langle \theta_h, a\rangle| \le m$, the accuracy restriction and $\zeta < 1/2$ give

$$
mu = \frac{128\epsilon}{H^2} \le \frac{128\zeta H}{384H^2} = \frac{\zeta}{3H} < \frac{1}{6H}. \tag{82}
$$

Thus, uniformly over stages and actions,

$$
\frac{1}{3H} = p_0 - \frac{1}{6H} \le P_h(x_G \mid x, a) \le p_0 + \frac{1}{6H} = \frac{2}{3H}. \tag{83}
$$

The remaining probability is assigned to x, so the transition kernels are well-defined.

Next, we verify the d-dimensional linear representation. Let

$$
\boldsymbol{\phi}(x,a) = \frac{(1, a^\top, 0)^\top}{\sqrt{m+1}}, \qquad \boldsymbol{\phi}(x_G,a) = e_d,
$$

and define the vector-valued transition measures by

$$
\boldsymbol{\nu}_h(x) = \sqrt{m+1}\,(1 - p_0, -u\theta_h^\top, 0)^\top,
$$

$$\boldsymbol{\nu}_h(x_G) = (\sqrt{m+1}\, p_0, \sqrt{m+1}\, u\theta_h^\top, 1)^\top.$$

Their dot products recover all four transition cases:

$$\langle \boldsymbol{\phi}(x,a), \boldsymbol{\nu}_h(x_G)\rangle = p_0 + u\langle \theta_h, a\rangle, \qquad \langle \boldsymbol{\phi}(x,a), \boldsymbol{\nu}_h(x)\rangle = 1 - p_0 - u\langle \theta_h, a\rangle,$$
$$\langle \boldsymbol{\phi}(x_G,a), \boldsymbol{\nu}_h(x_G)\rangle = 1, \qquad \langle \boldsymbol{\phi}(x_G,a), \boldsymbol{\nu}_h(x)\rangle = 0.$$

The reward parameter is e_d, since $\langle \boldsymbol{\phi}(x,a), e_d\rangle = 0$ and $\langle \boldsymbol{\phi}(x_G,a), e_d\rangle = 1$. Furthermore,

$$\|\boldsymbol{\phi}(x,a)\|_2^2 = \frac{1+\|a\|_2^2}{m+1} = 1, \qquad \|\boldsymbol{\phi}(x_G,a)\|_2 = \|e_d\|_2 = 1,$$
$$\boldsymbol{\nu}_h(\mathcal{S}) = (\sqrt{m+1}, 0, \ldots, 0, 1)^\top, \qquad \|\boldsymbol{\nu}_h(\mathcal{S})\|_2 = \sqrt{m+2} = \sqrt{d}.$$

This verifies the stated feature and parameter normalization.

Finally, let M_{ref} have the same rewards and features but no perturbation, that is, $u = 0$, and supply the common reference $Q_{\mathrm{ref}} = Q^{\star, M_{\mathrm{ref}}}$ to the learner. For each fixed accuracy ϵ, take the finite family indexed by

$$\Theta := \big(\{-1,+1\}^m\big)^H, \qquad |\Theta| = 2^{mH},$$

with the uniform prior on the sign arrays $\theta = (\theta_1, \ldots, \theta_H)$. The corresponding family is denoted by $\mathcal{M}_\zeta$, suppressing its dependence on ϵ. The supplied reference is the same for every θ, so it cannot reveal the unknown signs. Lemma D.1 verifies that all these models satisfy the required reference-closeness condition.

D.2 Supporting Lemmas

Lemma D.1 (Reference closeness)**.** *Every MDP in this construction satisfies*

$$\max_{h,s,a} |Q_{\mathrm{ref};h}(s,a) - Q_h^\star(s,a)| \le 64\epsilon \le \zeta H.$$

Proof. Let $V_{\mathrm{ref},h} := V_h^{\star, M_{\mathrm{ref}}}$, let $P_{\mathrm{ref},h}$ be the reference transition kernel, and fix a model M in the constructed family. Define

$$\delta_h := \max_{s,a} |Q_{\mathrm{ref};h}(s,a) - Q_h^\star(s,a;M)|, \qquad \delta_{H+1} := 0.$$

We suppress M from the notation for its optimal values. The rewards of the reference and target models are identical. Adding and subtracting $P_{\mathrm{ref},h} V_{h+1}^\star$ in their Bellman equations gives

$$\begin{aligned} &|Q_{\mathrm{ref};h}(s,a) - Q_h^\star(s,a)| \\ &\quad = |[P_{\mathrm{ref},h} V_{\mathrm{ref},h+1}](s,a) - [P_h V_{h+1}^\star](s,a)| \\ &\quad \le |[P_{\mathrm{ref},h}(V_{\mathrm{ref},h+1} - V_{h+1}^\star)](s,a)| + |[(P_{\mathrm{ref},h} - P_h) V_{h+1}^\star](s,a)|. \end{aligned} \tag{84}$$

For the first term, the transition probabilities sum to one, and the maximum over actions is 1-Lipschitz in the sup norm. Thus

$$\begin{aligned} |[P_{\mathrm{ref},h}(V_{\mathrm{ref},h+1} - V_{h+1}^\star)](s,a)| &\le \max_{s'} |V_{\mathrm{ref},h+1}(s') - V_{h+1}^\star(s')| \\ &\le \max_{s',a'} |Q_{\mathrm{ref};h+1}(s',a') - Q_{h+1}^\star(s',a')| = \delta_{h+1}. \end{aligned} \tag{85}$$

For the second term, the kernels coincide at x_G. At x, the construction gives the exact identity

$$[(P_{\mathrm{ref},h} - P_h) V_{h+1}^\star](x,a) = u\langle \theta_h, a\rangle \big(V_{h+1}^\star(x) - V_{h+1}^\star(x_G)\big).$$

Since $|\langle \theta_h, a\rangle| \le m$ and both continuation values are in $[0, H-h]$, this implies

$$|[(P_{\mathrm{ref},h} - P_h) V_{h+1}^\star](s,a)| \le mu(H-h) \qquad \text{for either state } s. \tag{86}$$

Combining (84)– (86) and maximizing over (s,a), we obtain

$$\delta_h \le \delta_{h+1} + mu(H-h).$$

Unrolling this recursion through the terminal condition gives

$$\begin{aligned}\delta_h &\le mu\sum_{j=h}^{H}(H-j) = mu\,\frac{(H-h)(H-h+1)}{2}\\ &\le mu\,\frac{H(H-1)}{2} = \frac{128\epsilon}{H^2}\frac{H(H-1)}{2}\\ &= 64\epsilon\frac{H-1}{H} \le 64\epsilon \le \frac{\zeta H}{6} \le \zeta H.\end{aligned}$$

The first equality sums the arithmetic progression; the second line uses (81); and the penultimate comparison uses $\epsilon \le \zeta H/384$. This proves the uniform closeness bound for every sign array in the family. □

Lemma D.2 (Decomposition)**.** *Let $L = \lfloor H/2 \rfloor$. For each $h \le L$, the probability of reaching x at stage h is at least $1/3$ under every policy, and*

$$V^\star_{h+1}(x_G) - V^\star_{h+1}(x) \ge H/6.$$

If $N_{h,j}(\theta)$ counts episodes in which the learner visits x at stage h and takes an action whose jth sign differs from $\theta_{h,j}$, then

$$\mathbb{E}_\theta[\mathrm{Regret}(K)] \ge \frac{Hu}{3}\sum_{h=1}^{L}\sum_{j=1}^{m}\mathbb{E}_\theta[N_{h,j}(\theta)].$$

Proof. Fix a sign array θ. We derive a lower bound on visits to x, a lower bound on the value gained by entering x_G, and then the resulting regret decomposition.

By (83), under any chosen action,

$$P_j(x \mid x, a) \ge 1 - \frac{2}{3H}.$$

This bound also holds after conditioning on any previous interaction history. Starting from x and applying the tower property at each stage,

$$\begin{aligned}\Pr_\theta(s_h = x) &\ge \left(1 - \frac{2}{3H}\right)^{h-1}\\ &\ge \left(1 - \frac{2}{3H}\right)^{H} \ge 1 - H\frac{2}{3H} = \frac{1}{3}.\end{aligned} \tag{87}$$

The second inequality uses $h - 1 \le H$ and a base in $(0, 1)$; the last is Bernoulli's inequality $(1-t)^H \ge 1 - Ht$. In particular, the result holds conditionally on the training history preceding each episode, even when the learner is randomized or history-dependent.

At stage h, the action $a = \theta_h$ maximizes the probability of moving to x_G, since $\max_a \langle \theta_h, a\rangle = m$. It is optimal because entering x_G cannot decrease future rewards. The optimal transition probability is the same at every stage: $q := p_0 + mu \le 2/(3H)$. Let $g_h := V^\star_h(x_G) - V^\star_h(x)$ and $g_{H+1} = 0$. The Bellman equations give

$$\begin{aligned}g_h &= \big(1 + V^\star_{h+1}(x_G)\big) - \big(qV^\star_{h+1}(x_G) + (1-q)V^\star_{h+1}(x)\big)\\ &= 1 + (1-q)g_{h+1}.\end{aligned}$$

Backward substitution yields

$$g_{h+1} = 1 + (1-q) + \cdots + (1-q)^{H-h-1} = \sum_{t=0}^{H-h-1}(1-q)^t. \tag{88}$$

For $h \le L := \lfloor H/2 \rfloor$, the number of summands is at least $H/2$. Hence

$$\begin{aligned}g_{h+1} &\ge (H-h)(1-q)^{H-h-1}\\ &\ge (H-h)\left(1 - \frac{2}{3H}\right)^{H}\end{aligned}$$

$$\geq \frac{H-h}{3} \geq \frac{H}{6}. \tag{89}$$

The first inequality bounds every summand by the smallest one; the second uses $q \leq 2/(3H)$; the third uses the same Bernoulli inequality as (87); and the last uses $h \leq H/2$.

Write $p_h(a) = p_0 + u\langle \theta_h, a\rangle$. The optimal and action-conditioned Bellman equations at x give

$$\begin{aligned} V_h^\star(x) - Q_h^\star(x,a) &= \big(q - p_h(a)\big)\big(V_{h+1}^\star(x_G) - V_{h+1}^\star(x)\big) \\ &= u\big(m - \langle \theta_h, a\rangle\big) g_{h+1} \\ &= 2u \sum_{j=1}^{m} \mathbb{1}\{a_j \neq \theta_{h,j}\}\, g_{h+1} \\ &\geq \frac{Hu}{3} \sum_{j=1}^{m} \mathbb{1}\{a_j \neq \theta_{h,j}\}, \qquad h \leq L. \end{aligned} \tag{90}$$

For the third line, each coordinate satisfies $1 - \theta_{h,j} a_j = 2\mathbb{1}\{a_j \neq \theta_{h,j}\}$. The last line uses (89). At x_G, every action has zero disadvantage.

To spell out the performance-difference identity, define $A_h^\star(s,a) := V_h^\star(s) - Q_h^\star(s,a) \geq 0$. For any learner and any episode, conditional on its initial history,

$$\begin{aligned} \mathbb{E}_\theta\left[\sum_{h=1}^{H} A_h^\star(s_h,a_h)\right] &= \mathbb{E}_\theta\left[\sum_{h=1}^{H} \big(V_h^\star(s_h) - r_h(s_h,a_h) - [P_h V_{h+1}^\star](s_h,a_h)\big)\right] \\ &= \mathbb{E}_\theta\left[\sum_{h=1}^{H} \big(V_h^\star(s_h) - r_h(s_h,a_h) - V_{h+1}^\star(s_{h+1})\big)\right] \\ &= V_1^\star(x) - \mathbb{E}_\theta\left[\sum_{h=1}^{H} r_h(s_h,a_h)\right]. \end{aligned}$$

The second equality is the tower property for the transition kernel; the third telescopes and uses $V_{H+1}^\star = 0$. Consequently, after summing over episodes and taking expectation over training randomness,

$$\begin{aligned} \mathbb{E}_\theta[\mathrm{Regret}(K)] &= \mathbb{E}_\theta\left[\sum_{k=1}^{K}\sum_{h=1}^{H} A_h^\star(s_h^k, a_h^k)\right] \\ &\geq \frac{Hu}{3} \sum_{h=1}^{L}\sum_{j=1}^{m} \mathbb{E}_\theta\left[\sum_{k=1}^{K} \mathbb{1}\{s_h^k = x,\ a_{h,j}^k \neq \theta_{h,j}\}\right] \\ &= \frac{Hu}{3} \sum_{h=1}^{L}\sum_{j=1}^{m} \mathbb{E}_\theta[N_{h,j}(\theta)]. \end{aligned}$$

Here we discard only nonnegative disadvantages at later stages and use (90) at the earlier stages. The count in brackets is precisely $N_{h,j}(\theta)$ in the lemma. This proves the claimed decomposition, including for history-dependent learners. □

D.3 Detailed Proof

Proof of Theorem 4.1. Let $\Theta = (\{-1,+1\}^m)^H$ be the sign-array family in Section D.1. For fixed $\theta \in \Theta$, denote by $\mathbb{P}_\theta$ the law of the complete K-episode interaction history, including states, actions, rewards, and learner randomization. We show that neighboring sign arrays are statistically difficult to distinguish, and then use Lemma D.2 to convert sign errors into regret.

Fix a coordinate (h,j) and let θ' differ from θ only in that coordinate. The common reference Q_{ref} is identical in the two models. Conditional on a realized history, the learner's action rule is also the same in the two models. The rewards coincide, and the transition kernels differ only at stage h when

the state is x. By the chain rule of relative entropy, all other conditional contributions vanish, leaving

$$\mathrm{KL}(\mathbb{P}_\theta \,\|\, \mathbb{P}_{\theta'}) = \sum_{k=1}^{K} \mathbb{E}_\theta \Big[\mathbb{1}\{s_h^k = x\} \mathrm{kl}\big(p_h^\theta(a_h^k), p_h^{\theta'}(a_h^k)\big) \Big], \tag{91}$$

where kl denotes Bernoulli relative entropy and $p_h^\theta(a) = p_0 + u\langle \theta_h, a \rangle$. The changed coordinate gives

$$|p_h^\theta(a) - p_h^{\theta'}(a)| = |2u\theta_{h,j} a_j| = 2u.$$

For $p, q \in (0,1)$, the elementary inequality $\log t \le t - 1$ implies

$$\begin{aligned} \mathrm{KL}(p,q) &= p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \\ &\le p\left(\frac{p}{q} - 1\right) + (1-p)\left(\frac{1-p}{1-q} - 1\right) \\ &= \frac{(p-q)^2}{q(1-q)}. \end{aligned} \tag{92}$$

By (83), both parameters lie in $[1/(3H), 2/(3H)]$. As $H \ge 3$,

$$q \ge \frac{1}{3H}, \qquad 1 - q \ge 1 - \frac{2}{3H} \ge \frac{3}{4}, \qquad q(1-q) \ge \frac{1}{4H}.$$

Substituting into (92) and then (91) yields

$$\begin{aligned} \mathrm{KL}\big(p_h^\theta(a), p_h^{\theta'}(a)\big) &\le \frac{(2u)^2}{1/(4H)} = 16Hu^2, \\ \mathrm{KL}(\mathbb{P}_\theta \,\|\, \mathbb{P}_{\theta'}) &\le 16Hu^2 \sum_{k=1}^{K} \mathbb{E}_\theta[\mathbb{1}\{s_h^k = x\}] \le 16HKu^2. \end{aligned} \tag{93}$$

This bound accounts for all online observations, not only reward observations or a single stage viewed in isolation.

Suppose

$$K \le \frac{1}{288Hu^2} = \frac{m^2 H^3}{288 \cdot 128^2 \epsilon^2}. \tag{94}$$

Pinsker's inequality (Lattimore and Szepesvári, 2020) and (93) imply

$$\begin{aligned} \mathrm{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) &\le \sqrt{\frac{1}{2}\mathrm{KL}(\mathbb{P}_\theta \,\|\, \mathbb{P}_{\theta'})} \le \sqrt{8HKu^2} \\ &\le \sqrt{\frac{8}{288}} = \frac{1}{6}. \end{aligned} \tag{95}$$

The equality in (94) substitutes $u = 128\epsilon/(mH^2)$ from (81).

For a fixed complete history ω, let $T_h(\omega) := \sum_{k=1}^{K} \mathbb{1}\{s_h^k = x\}$. The arrays θ and θ' prescribe opposite signs at (h, j). At every visit to x, exactly one of these two prescriptions disagrees with the observed sign $a_{h,j}^k$. Hence, as functions of the same history,

$$N_{h,j}(\theta; \omega) + N_{h,j}(\theta'; \omega) = T_h(\omega). \tag{96}$$

All three counts lie in $[0, K]$. For any history-measurable function $F \in [0, K]$, the layer-cake identity and the definition of total variation give

$$\begin{aligned} |\mathbb{E}_\theta F - \mathbb{E}_{\theta'} F| &= \left| \int_0^K \big(\mathbb{P}_\theta(F > t) - \mathbb{P}_{\theta'}(F > t)\big)\, \mathrm{d}t \right| \\ &\le K\, \mathrm{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}). \end{aligned}$$

Apply this to $F = N_{h,j}(\theta'; \cdot)$, then use (96):

$$\mathbb{E}_\theta N_{h,j}(\theta) + \mathbb{E}_{\theta'} N_{h,j}(\theta')$$

$$
\begin{aligned}
&= \mathbb{E}_\theta\big[N_{h,j}(\theta) + N_{h,j}(\theta')\big] + (\mathbb{E}_{\theta'} - \mathbb{E}_\theta)N_{h,j}(\theta') \\
&\geq \mathbb{E}_\theta T_h - K\,\mathrm{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \\
&\geq \frac{K}{3} - \frac{K}{6} = \frac{K}{6}.
\end{aligned} \tag{97}
$$

The last line uses the reach-probability conclusion of Lemma D.2 for each episode and (95).

The map that flips coordinate (h, j) is a bijection of Θ and its own inverse. Summing (97) over all $\theta \in \Theta$ therefore gives

$$
\begin{aligned}
2\sum_{\theta\in\Theta} \mathbb{E}_\theta N_{h,j}(\theta) &= \sum_{\theta\in\Theta} \left(\mathbb{E}_\theta N_{h,j}(\theta) + \mathbb{E}_{\theta'} N_{h,j}(\theta')\right) \\
&\geq |\Theta|\frac{K}{6}.
\end{aligned}
$$

Dividing by $2|\Theta|$ yields, for every coordinate,

$$
\mathbb{E}_{\theta\sim\mathrm{Unif}(\Theta)}\mathbb{E}_\theta N_{h,j}(\theta) \geq \frac{K}{12}. \tag{98}
$$

Now average the regret decomposition in Lemma D.2 over the same prior:

$$
\begin{aligned}
\mathbb{E}_{\theta\sim\mathrm{Unif}(\Theta)}\mathbb{E}_\theta[\mathrm{Regret}(K)] &\geq \frac{Hu}{3}\sum_{h=1}^{L}\sum_{j=1}^{m} \mathbb{E}_{\theta\sim\mathrm{Unif}(\Theta)}\mathbb{E}_\theta N_{h,j}(\theta) \\
&\geq \frac{Hu}{3}\, Lm\frac{K}{12} = \frac{HuLmK}{36} \\
&\geq \frac{umH^2K}{108} = \frac{128}{108}\epsilon K > \epsilon K.
\end{aligned} \tag{99}
$$

The second line uses (98); the third uses $L = \lfloor H/2 \rfloor \geq H/3$ for integer $H \geq 3$; and the final equality substitutes $u = 128\epsilon/(mH^2)$.

Equation (99) contradicts expected average regret at most ϵ whenever the budget satisfies (94). Hence the required number of episodes is at least a constant multiple of m^2H^3/ϵ^2. Since $m = d - 2 \geq d/3$ for $d \geq 3$, this is $\Omega(H^3d^2/\epsilon^2)$. Lemma D.1 verifies $\|Q_{\mathrm{ref}} - Q^{\star,M}\|_\infty \leq \zeta H$ for every member of the same family, and the construction supplies an identical reference to every learner. Taking $c_0 = 1/384$ establishes all claims in Theorem 4.1. □

E Technical Lemmas

Lemma E.1 (Hoeffding's Inequality). *Let $X_1, \ldots, X_n$ be independent random variables such that $a_i \leq X_i \leq b_i$ almost surely for all $i \in [n]$, and define $S_n := \sum_{i=1}^n X_i$ with $\mu := \mathbb{E}[S_n]$. Then, with probability at least $1 - \delta$, we have*

$$
S_n - \mu \leq \sqrt{\frac{1}{2}\Big(\sum_{i=1}^{n}(b_i - a_i)^2\Big)\log\frac{1}{\delta}}.
$$

Lemma E.2 (Azuma–Hoeffding Inequality). *Let $\{X_t\}_{t=0}^T$ be a martingale adapted to a filtration $\{\mathcal{F}_t\}_{t=0}^T$, and define the martingale difference sequence $D_t = X_t - X_{t-1}$. Assume that the increments are almost surely bounded:*

$$
|D_t| \leq c_t \quad \text{for all } t = 1, \ldots, T.
$$

Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$
X_T - X_0 \;\leq\; \sqrt{2\log\left(\frac{1}{\delta}\right)\sum_{t=1}^{T} c_t^2}.
$$

Lemma E.3 (Nonnegative predictable costs)**.** *Let $0 \le X_i \le B$ be adapted observations and let $m_i = \mathbb{E}[X_i \mid \mathcal{G}_{i-1}]$ for their pre-observation filtration. For each $\eta \in (0,1)$, each of the following holds with probability at least $1-\eta$:*

$$\sum_i m_i \le 2\sum_i X_i + 2B\log(1/\eta), \qquad \sum_i X_i \le 2\sum_i m_i + B\log(1/\eta).$$

The same statements apply to a predictable selection of nonnegative costs by including the selection indicator in X_i.

Proof. If $B = 0$, every observation and conditional mean is zero, so both inequalities hold. Assume $B > 0$ and write

$$U_i := X_i/B \in [0,1], \qquad \mu_i := \mathbb{E}[U_i \mid \mathcal{G}_{i-1}] = m_i/B.$$

We prove the two one-sided inequalities separately. Each has failure probability at most η; simultaneous use of both may be obtained by splitting a failure budget between them.

The graph of a convex function lies below the chord connecting two points on its graph. Applying this to $u \mapsto e^{-u}$ on $[0,1]$ gives

$$e^{-u} \le (1-u)e^0 + ue^{-1} = 1 - (1-e^{-1})u.$$

Let $a := 1 - e^{-1}$. Taking conditional expectations and using $1 - v \le e^{-v}$ yields

$$\begin{aligned} \mathbb{E}[e^{-U_i} \mid \mathcal{G}_{i-1}] &\le 1 - a\mu_i \le e^{-a\mu_i}, \\ \mathbb{E}[e^{a\mu_i - U_i} \mid \mathcal{G}_{i-1}] &\le 1. \end{aligned} \tag{100}$$

Since μ_i is $\mathcal{G}_{i-1}$-measurable, the process

$$Z_n^- := \exp\left(a\sum_{i=1}^n \mu_i - \sum_{i=1}^n U_i\right), \qquad Z_0^- = 1,$$

is a nonnegative supermartingale: indeed,

$$\mathbb{E}[Z_n^- \mid \mathcal{G}_{n-1}] = Z_{n-1}^- \mathbb{E}[e^{a\mu_n - U_n} \mid \mathcal{G}_{n-1}] \le Z_{n-1}^-.$$

Thus $\mathbb{E}Z_n^- \le 1$. By Markov's inequality, $\Pr(Z_n^- > 1/\eta) \le \eta$. On the complementary event,

$$\begin{aligned} a\sum_{i=1}^n \mu_i - \sum_{i=1}^n U_i &\le \log(1/\eta), \\ \sum_{i=1}^n m_i &\le \frac{1}{a}\sum_{i=1}^n X_i + \frac{B}{a}\log(1/\eta) \\ &\le 2\sum_{i=1}^n X_i + 2B\log(1/\eta), \end{aligned}$$

where $a^{-1} = (1-e^{-1})^{-1} < 2$. This proves the first inequality.

The chord bound for $u \mapsto e^u$ gives $e^u \le 1 + (e-1)u$ for $u \in [0,1]$. Let $b := e - 1$. Then

$$\begin{aligned} \mathbb{E}[e^{U_i} \mid \mathcal{G}_{i-1}] &\le 1 + b\mu_i \le e^{b\mu_i}, \\ \mathbb{E}[e^{U_i - b\mu_i} \mid \mathcal{G}_{i-1}] &\le 1. \end{aligned}$$

The same tower-property argument shows that

$$Z_n^+ := \exp\left(\sum_{i=1}^n U_i - b\sum_{i=1}^n \mu_i\right)$$

is a nonnegative supermartingale with initial value one. Except on an event of probability η, Markov's inequality yields

$$\sum_{i=1}^n U_i - b\sum_{i=1}^n \mu_i \le \log(1/\eta),$$

$$\sum_{i=1}^{n} X_i \le b \sum_{i=1}^{n} m_i + B \log(1/\eta)$$
$$\le 2 \sum_{i=1}^{n} m_i + B \log(1/\eta),$$

since $b = e - 1 < 2$. This proves the second inequality.

Let $J_i \in \{0, 1\}$ be $\mathcal{G}_{i-1}$-measurable and define $X_i' := J_i X_i$. Then $0 \le X_i' \le B$ and

$$\mathbb{E}[X_i' \mid \mathcal{G}_{i-1}] = J_i \mathbb{E}[X_i \mid \mathcal{G}_{i-1}] = J_i m_i.$$

Applying either of the preceding arguments to X_i' proves the corresponding selected-cost inequality. Predictability is essential to this last equality; selection after observing X_i is not covered by this assertion. □

F Verification for Example 5.4

We verify Example 5.4 through an explicit construction. Let $H \ge 2$, $0 < \beta \le b \le 1$, $\rho \in [0, 1]$, and $p := \rho/H$. Consider a base linear MDP with state space $\mathcal{X}$, finite action space $\mathcal{A}$, transition kernel

$$\bar{P}_h(\cdot \mid x, a) = \left\langle \bar{\boldsymbol{\phi}}(x, a), \bar{\boldsymbol{\mu}}_h(\cdot) \right\rangle,$$

and reward

$$\bar{r}_h(x, a) = \left\langle \bar{\boldsymbol{\phi}}(x, a), \boldsymbol{\theta}_h \right\rangle \in [0, 1 - b].$$

Construction of the misspecification. Augment the state space to $\mathcal{S} := \mathcal{X} \times \{0, 1\}$, where $s = (x, c)$ and c is a corruption indicator. Let ν_p be the Bernoulli distribution with $\nu_p(\{1\}) = p$. Define the transition kernel by

$$P_h(\mathrm{d}x', \mathrm{d}c' \mid (x, c), a) = \bar{P}_h(\mathrm{d}x' \mid x, a)\, \nu_p(\mathrm{d}c').$$

Thus, the next corruption indicator is sampled independently of the current state and action.

The augmented transition kernel retains the linear representation with

$$\boldsymbol{\phi}((x, c), a) := \bar{\boldsymbol{\phi}}(x, a),$$
$$\boldsymbol{\mu}_h(\mathrm{d}x', \mathrm{d}c') := \bar{\boldsymbol{\mu}}_h(\mathrm{d}x')\, \nu_p(\mathrm{d}c').$$

In particular,

$$P_h(\cdot \mid s, a) = \left\langle \boldsymbol{\phi}(s, a), \boldsymbol{\mu}_h(\cdot) \right\rangle.$$

For each $h \in \{2, \dots, H\}$, choose a function $g_h : \mathcal{A} \to [\beta, b]$ and define

$$\xi_h((x, c), a) := \begin{cases} 0, & h = 1, \\ c\, g_h(a), & 2 \le h \le H. \end{cases}$$

The reward is

$$r_h((x, c), a) := \bar{r}_h(x, a) + \xi_h((x, c), a) \in [0, 1].$$

For $h \ge 2$, this residual is not representable using $\boldsymbol{\phi}$: the states $(x, 0)$ and $(x, 1)$ have identical features for every action, whereas their residuals differ by $g_h(a) \ge \beta$.

Idealized reference and separation. Let $V^\star$ be the optimal value function of the augmented MDP, with $V_{H+1}^\star \equiv 0$. For this existence construction, define the idealized linear reference

$$\boldsymbol{w}_h := \boldsymbol{\theta}_h + \int_{\mathcal{S}} V_{h+1}^\star(s')\, \mathrm{d}\boldsymbol{\mu}_h(s'),$$
$$Q_{\mathrm{ref};h}(s, a) := \left\langle \boldsymbol{\phi}(s, a), \boldsymbol{w}_h \right\rangle.$$

This definition uses the true continuation value $V^\star_{h+1}$; it does not assert that the reference can be computed from the linear model components alone.

By the Bellman optimality equation,

$$\begin{aligned} Q^\star_h(s,a) &= r_h(s,a) + [P_h V^\star_{h+1}](s,a) \\ &= \xi_h(s,a) + \left\langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h + \int_{\mathcal{S}} V^\star_{h+1}(s')\, \mathrm{d}\boldsymbol{\mu}_h(s') \right\rangle \\ &= \xi_h(s,a) + Q_{\mathrm{ref};h}(s,a). \end{aligned}$$

Define the corrupted subsets by

$$\mathcal{S}'_1 := \varnothing, \qquad \mathcal{S}'_h := \mathcal{X} \times \{1\}, \qquad 2 \le h \le H.$$

The reference is exact outside $\mathcal{S}'_h \times \mathcal{A}$, while

$$|Q^\star_h(s,a) - Q_{\mathrm{ref};h}(s,a)| = \xi_h(s,a) \ge \beta$$

on $\mathcal{S}'_h \times \mathcal{A}$. Hence Q_{ref} satisfies Definition 5.1.

Policy-uniform coverage. For every $h \in \{2, \dots, H\}$ and every (s,a), the transition construction gives

$$P_{h-1}(\mathcal{S}'_h \mid s,a) = p.$$

Consequently, under any policy π and any initial-state distribution,

$$\Pr_\pi(s_h \in \mathcal{S}'_h) = p, \qquad 2 \le h \le H.$$

Since $\xi_1 \equiv 0$, the reference is exact at the first stage. Therefore,

$$\begin{aligned} &\mathbb{E}_{(s,a)\sim d^\pi_h}\left[\mathbf{1}\left\{Q^\star_h(s,a) \ne Q_{\mathrm{ref};h}(s,a)\right\}\right] \\ &\quad = \begin{cases} 0, & h = 1, \\ \rho/H, & 2 \le h \le H, \end{cases} \le \rho. \end{aligned}$$

This verifies Assumption 5.3. In fact, the construction also satisfies the stronger bound

$$\sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d^\pi_h}\left[\mathbf{1}\left\{Q^\star_h(s,a) \ne Q_{\mathrm{ref};h}(s,a)\right\}\right] = \frac{(H-1)\rho}{H} \le \rho.$$

Reference misspecification versus online realizability. The misspecification above concerns the reference feature map $\boldsymbol{\phi}$, not necessarily the function class used for online learning. To satisfy Assumption 6.1, the online class must also represent the reward residual. For this construction, this can be achieved by using the richer feature map

$$\boldsymbol{\psi}((x,c),a) := \left(\bar{\boldsymbol{\phi}}(x,a), c\,\mathbf{e}_a\right),$$

where $\mathbf{e}_a$ is the canonical basis vector indexed by $a \in \mathcal{A}$. Indeed, the reward and every Bellman backup $r_h + P_h V$ are linear in $\boldsymbol{\psi}$. With sufficiently large coefficient and output bounds, the corresponding online class therefore contains the required Bellman backups, even though the reference uses only $\boldsymbol{\phi}$.

G Empirical Results

G.1 MuJoCo Results

Results and Analysis. We conduct an empirical study of O2O-LSVI on standard MuJoCo AntMaze environments, leveraging D4RL datasets (Fu et al., 2020) for offline RL pretraining. In the offline phase, we adopt the Cal-QL (Nakamoto et al., 2023) pretraining strategy and use an ensemble of Q-networks. During online adaptation, we instantiate the confidence interval in the inclusion test (4) using the ensemble standard deviation as the interval width. We compare our method against direct

adaptation via Cal-QL (Nakamoto et al., 2023) and the offline RL baselines CQL (Kumar et al., 2020) and IQL (Kostrikov et al., 2021). As shown in Table 1, our method matches or outperforms these baselines. These final-score comparisons do not isolate online sample efficiency, since the reported offline starting scores differ.

Settings	Offline RL		Offline-to-Online Adaptation	
	CQL	IQL	Cal-QL	Ours
Umaze	94.0±1.6	77.0±0.7	76.8±7.5→99.8±0.4	85.8±3.3→99.8±0.4
Medium-Play	59.0±11.2	71.8±3.0	71.8±3.3→98.8±1.6	70.3±2.0→99.3±1.3
Large-Play	28.8±7.8	38.5±8.7	31.8±8.9→97.3±1.8	35.3±4.0→98.5±1.6

Table 1: **Empirical Results on AntMaze.** We compare our method with existing offline-to-online adaptation and offline RL baselines on AntMaze. Our method achieves performance comparable to or better than existing baselines. Results are reported as D4RL scores averaged over four random seeds.

Experimental Details. We build our method upon the Cal-QL offline pretraining procedure (Nakamoto et al., 2023). During online fine tuning, we form the confidence interval

$$[\mathrm{mean}(Q_{\mathrm{ens}}) - \mathrm{std}(Q_{\mathrm{ens}}),\ \mathrm{mean}(Q_{\mathrm{ens}}) + \mathrm{std}(Q_{\mathrm{ens}})]$$

from a Q-function ensemble, and use the resulting uncertainty estimate to regulate the Q-function during the SAC policy update (Haarnoja et al., 2018). The reported experiments use numerical reference-switching implementations, not an explicit enumeration of the finite covers in the theoretical algorithm. The finite-cover and projection safeguards in the revised analysis are not a claim that the original experimental runs were re-executed with these changes. Ensemble intervals are heuristic uncertainty estimates rather than the proved confidence sets. The hyperparameter settings are provided in Table 2. We implement our method based on the CORL repository (Tarasov et al., 2023), which also serves as the codebase for the offline RL baselines CQL (Kumar et al., 2020), IQL (Kostrikov et al., 2021), and Cal-QL (Nakamoto et al., 2023).

Hyperparameter	Value
Ensemble Size	5
β	50
CQL Conservative Coefficient α (Offline & Online)	5.0
Discount factor	0.99
Q Network Learning Rate	3e-4
Policy Network Learning Rate	1e-4
Batch Size	256
Soft Target Update Portion	0.005
Number of Layers for Q Function	5
Number of Layers for Policy Network	3
Hidden Dimension (Q Function and Policy)	256
CQL Temperature	1.0

Table 2: **Hyperparameters.** We demonstrate the hyperparameters used in the empirical experiment with the practical implementation of O2O-LSVI.

G.2 Experimental Result with Linear MDP

Experimental setup. We compare O2O-LSVI and LSVI-UCB in a linear MDP with 10 states, 2 actions, horizon $H = 3$, and initial state 0. Rewards are Bernoulli with mean

$$r_h(s,a) = \frac{1}{2} + \frac{\gamma}{2}\sigma_h(2a-1), \qquad \gamma = 0.10. \tag{101}$$

The stage signs $\sigma_h \in \{-1,+1\}$ are fixed within each run. Transitions follow stage-dependent, nonuniform distributions independent of the current state and action, giving a fixed action gap of 0.10 throughout the experiment.

We construct an oracle reference by subtracting 0.44 from the rewards at the three reachable suboptimal entries with lowest state occupancy, denoted by $\mathcal{D}$, and planning with the true transitions. This preserves the optimal policy and gives

$$V_h^{\mathrm{ref}}(s) = V_h^*(s), \qquad Q_h^*(s,a) - Q_h^{\mathrm{ref}}(s,a) = 0.44\,\mathbf{1}\{(h,s,a) \in \mathcal{D}\}. \tag{102}$$

The 60 entries comprise 30 exact optimal, 27 exact suboptimal, and 3 misspecified suboptimal references, giving a table-entry fraction of 0.05, not the occupancy coefficient ρ. Here the policy-uniform coefficient is $\rho = \max_h \sum_s p_h(s)\mathbb{1}\{\exists a : (h,s,a) \in \mathcal{D}\}$, where p_h is the stage-h state distribution. Its numerical value requires the actual transition probabilities and is not inferred from the entry count. Both learners use the same norm-bounded, $d = 5$ feature map: two base coordinates augmented with three sparse indicators. True rewards, reference rewards, and transitions remain exactly linear in this map. Only O2O-LSVI additionally receives Q^{ref}.

Both learners start without online samples and use stagewise ridge regression with $\lambda = 1$, remaining-horizon upper clipping, and the common confidence multiplier

$$\alpha_k = 0.15H\sqrt{d\log\left(\frac{2dH\max\{k,2\}}{0.05}\right)}. \tag{103}$$

The effective scale is $0.15 = 0.5 \times 0.3$. O2O-LSVI uses the reference only when its entire pessimistic–optimistic interval lies within $Q^{\mathrm{ref}} \pm 0.20$, otherwise retaining its optimistic estimate. Thus the containment parameter is $\beta_{\mathrm{ref}} = 0.40$, while the reference error is $1.10\beta_{\mathrm{ref}} = 0.44$.

We run $K = 10^6$ episodes per algorithm for each of five seeds (3×10^6 steps per algorithm per seed), using paired common random numbers and shared random tie-breaking. We evaluate cumulative pseudo-regret using exact dynamic programming,

$$R_k = \sum_{i=1}^{k}\left[V_1^*(0) - V_1^{\pi_i}(0)\right], \tag{104}$$

and report 92 saved logarithmic checkpoints per seed. Regret curves show means and pointwise 95% Student-t intervals with four degrees of freedom; differences are paired within seed before aggregation.

Results and interpretation. Figure 1(a) shows that O2O-LSVI achieves substantially lower final regret than LSVI-UCB, reducing the mean regret from 1145.2 to 966.4 (15.6%), with lower regret across all five seeds. As shown in Figure 1(b), this improvement emerges after an initial transient cost: O2O-LSVI initially incurs higher regret but eventually overtakes LSVI-UCB and accumulates regret more slowly thereafter.

Figure 1(c) further illustrates the behavior of the reference-switching mechanism. The algorithm eventually accepts all correctly specified reference entries while consistently rejecting the misspecified ones. Notably, exact but suboptimal entries are also retained, which can help reduce optimistic overestimation of competing actions.

Because the reference preserves an optimal greedy policy, a frozen-reference greedy baseline has zero pseudo-regret in this construction. Thus this experiment illustrates certification and reference-switching behavior relative to learning from scratch, not recovery from a suboptimal reference policy. Overall, these results demonstrate a finite-sample benefit of reference switching relative to LSVI-UCB in this fixed-gap setting. However, they should not be interpreted as evidence of an asymptotic improvement, since both regret curves continue to grow and the experiment considers only a small number of seeds and a single reference configuration.

H Limitations

The paper's theoretical analysis relies on Assumptions 6.1 and 5.3. When these assumptions do not hold, our analysis has limited ability to characterize the algorithm's behavior. Moreover, our algorithm is tailored to settings with a β-separable structure; without such structure, the proposed method may not enjoy a provable advantage over standard online RL.

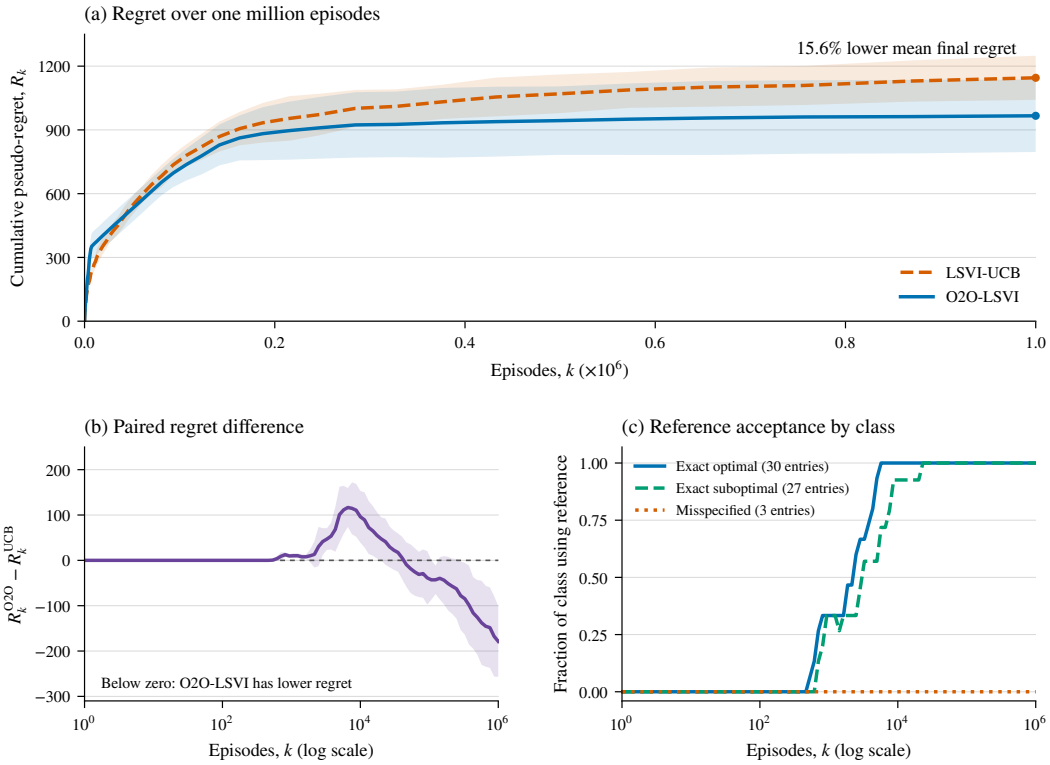

Figure 1: **Results on Linear MDP.** O2O-LSVI and LSVI-UCB with a fixed true action gap of 0.10, reference tolerance $\beta_{\text{ref}} = 0.40$, and reference error 0.44 on 3 of 60 entries. Both learners use $d = 5$ features. (a) Mean cumulative pseudo-regret across five seeds, with a linear episode axis. (b) Mean paired regret difference; negative values favor O2O-LSVI. Shaded regions in (a) and (b) are pointwise 95% Student-t intervals with four degrees of freedom, not simultaneous confidence bands. (c) Mean fraction of entries within each reference class satisfying the containment test at a checkpoint; these are table-entry fractions, not visitation-weighted or cumulative rates. Panels (b) and (c) use logarithmic episode axes. Lines connect saved checkpoints; panel (a) also includes the known origin $R_0 = 0$.