Actor-Critics Can Achieve Optimal Sample Efficiency

Kevin Tan [*1] **Wei Fan** [*1] **Yuting Wei** [1]

Abstract

Actor-critic algorithms have become a cornerstone in reinforcement learning (RL), leveraging the strengths of both policy-based and value-based methods. Despite recent progress in understanding their statistical efficiency, no existing work has successfully learned an ϵ-optimal policy with a sample complexity of $O(1/\epsilon^2)$ trajectories with general function approximation when strategic exploration is necessary. We address this open problem by introducing a novel actor-critic algorithm that attains a sample-complexity of $O(dH^5 \log|\mathcal{A}|/\epsilon^2 + dH^4 \log|\mathcal{F}|/\epsilon^2)$ trajectories, and accompanying $\sqrt{T}$ regret when the Bellman eluder dimension d does not increase with T at more than a $\log T$ rate. Here, $\mathcal{F}$ is the critic function class, and $\mathcal{A}$ is the action space. Our algorithm integrates optimism, off-policy critic estimation targeting the optimal Q-function, and rare-switching policy resets. We extend this to the setting of Hybrid RL, where we show that initializing the critic with offline data yields sample efficiency gains, and also provide a *non-optimistic* provably efficient actor-critic algorithm, addressing another open problem in the literature. Numerical experiments support our theoretical findings.

1. Introduction

Actor-critic algorithms have emerged as a foundational approach in reinforcement learning (RL), mitigating the deficiencies of both policy-based and value-based approaches (Sutton and Barto, 2018; Mnih et al., 2016; Haarnoja et al., 2018). These methods combine two components: an actor, which directly learns and improves the policy, and a critic, which evaluates the policy's quality. Given their popularity, significant recent progress has been made in understanding their theoretical underpinnings and statistical efficiency, especially in the presence of function approximation (Cai et al., 2024; Zhong and Zhang, 2023; Sherman et al., 2024; Liu et al., 2023b) – which is required in real-world applications with prohibitively large state and action spaces.

However, much existing work (Abbasi-Yadkori et al., 2019; Neu et al., 2017; Liu et al., 2023a; Bhandari and Russo, 2022; Agarwal et al., 2021; Cen et al., 2022; Gaur et al., 2024) on the convergence of actor-critic methods requires assumptions on the reachability of the state-action space or on the coverage of the sampled data. Liu et al. (2023b) remark that this implies that the state-space is either well-explored or easy to explore. This allows the agent to bypass the need to actively explore the state-action space, making learning significantly easier.[1] Therefore, these approaches analyze actor-critic methods from an optimization perspective and do not address the problem of exploration (Efroni et al., 2020) – a salient problem that we seek to tackle, hence the need for strategic exploration.

Without reachability assumptions, policy gradient methods struggle due to a lack of strategic exploration.[2] A recent line of work utilizes optimism to address this. Efroni et al. (2020); Wu et al. (2022) and Cai et al. (2024) achieve $\sqrt{T}$ regret within the settings of tabular and linear mixture MDPs respectively, with Wu et al. (2022) attaining the minimax-optimal rate. Still, Zhong and Zhang (2023) remark that these analyses do not generalize to more general MDPs due to the need to cover an exponentially growing policy class.

Within linear MDPs, Sherman et al. (2024) and Cassel and Rosenberg (2024) have very recently been able to obtain the optimal rate of $\sqrt{T}$ regret or $1/\epsilon^2$ sample complexity. They do so via methodological advancements (specific to linear MDPs) that let them overcome the growing policy class issue. However, the problem is unresolved with general function approximation – the best known algorithm from Liu et al. (2023b) requires at least $1/\epsilon^3$ samples, increasing to $1/\epsilon^4$ when the policy class grows exponentially.

[*]Equal contribution [1]Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, USA. Correspondence to: Kevin Tan <kevtan@wharton.upenn.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

[1]Under coverage/reachability assumptions, the linear convergence of policy-based methods (Lan, 2022; Xiao, 2022; Yuan et al., 2023) and the gradient domination lemma (Kumar et al., 2024; Mei et al., 2022) allow natural actor-critics to learn an ϵ-optimal policy within $1/\epsilon^2$ samples, although vanilla policy gradient methods can take (super) exponential time to converge (Li et al., 2021).

[2]To illustrate, this was not solved until Agarwal et al. (2020), who required $1/\epsilon^{11}$ samples to do so in linear MDPs.

Algorithm	Sample Complexity	Regret	Remarks
Agarwal et al. (2020)	$d^3H^{15}\log\lvert\mathcal{A}\rvert/\epsilon^{11}$	None	
Zanette et al. (2021)	$H^4\log\lvert\mathcal{A}\rvert/\epsilon^2 + d^3H^{13}\log\lvert\mathcal{A}\rvert/\epsilon^3$	$\sqrt{H^4\log\lvert\mathcal{A}\rvert\log\lvert\mathcal{A}\rvert T} + \sqrt{d^3H^{13}T}$	Linear MDPs only
Zhong and Zhang (2023)	$d^3H^8\log\lvert\mathcal{A}\rvert/\epsilon^4 + d^5H^4/\epsilon^2$	$\sqrt{d^3H^8\log\lvert\mathcal{A}\rvert T} + \sqrt{d^5H^4T}$	
Sherman et al. (2024)	$d^4H^7\log\lvert\mathcal{A}\rvert/\epsilon^2$	$\sqrt{d^4H^7\log\lvert\mathcal{A}\rvert T}$	
Cassel and Rosenberg (2024)	$dH^5\log\lvert\mathcal{A}\rvert/\epsilon^2 + d^3H^4/\epsilon^2$	$\sqrt{dH^5\log\lvert\mathcal{A}\rvert T} + \sqrt{d^3H^4T}$	
Zhou et al. (2023)	$(\log\lvert\mathcal{A}\rvert + C^3_{\text{npg}} \wedge C^6_{\text{off}})\log\lvert\mathcal{F}\rvert H^{14}/\epsilon^6$	Linear	Requires offline data
Liu et al. (2023b)	$d\log\lvert\mathcal{A}\rvert\log\lvert\mathcal{F}\rvert H^6/\epsilon^3$	None	"Good" case only, $1/\epsilon^4$ generally
DOUHUA (Algorithm 1)	$H^4\log\lvert\mathcal{A}\rvert/\epsilon^2 + dH^4\log\lvert\mathcal{F}\rvert/\epsilon^2$	$\sqrt{H^4\log\lvert\mathcal{A}\rvert T} + \sqrt{dH^4\log\lvert\mathcal{F}\rvert T}$	"Good" case only, vacuous generally
NORA (Algorithm 2)	$dH^5\log\lvert\mathcal{A}\rvert/\epsilon^2 + dH^4\log\lvert\mathcal{F}\rvert/\epsilon^2$	$\sqrt{dH^5\log\lvert\mathcal{A}\rvert T} + \sqrt{dH^4\log\lvert\mathcal{F}\rvert T}$	Holds generally

Table 1. Comparison of our work to existing literature. Algorithm 2 achieves the best known bound for actor-critic methods with general function approximation, and resolves an open problem on whether $\sqrt{T}$ regret or $1/\epsilon^2$ sample complexity is achievable in this scenario.

An open problem. *No actor-critic algorithm with general function approximation is currently known to achieve $1/\epsilon^2$ sample complexity or $\sqrt{T}$ regret* in this more challenging setting where strategic exploration is necessary. Zhong and Zhang (2023) and Liu et al. (2023b) remark that a way forward to achieve the desired $1/\epsilon^2$ sample complexity remains unclear, and raise the open problem:

Can actor-critic or policy optimization algorithms achieve $1/\epsilon^2$ sample complexity or $\sqrt{T}$ regret with general function approximation and when strategic exploration is necessary?

This paper. We resolve this open problem in the affirmative. As a warm-up, we first consider an easy case – where the complexity of the class of policies considered by the policy optimization procedure does not increase exponentially with the number of (critic) updates.[3] Then, a simple modification to the GOLF algorithm of Jin et al. (2021) allows one to achieve a regret of $\sqrt{H^4\log|\mathcal{A}|T} + \sqrt{dH^4\log|\mathcal{F}|\,T}$, in line with the results of (Efroni et al., 2020; Cai et al., 2024) for tabular and linear mixture MDPs respectively.

However, this is not the case in many practical scenarios – for example, where one uses linear models, decision trees, neural networks, or even random forests for the critic. In this much harder setting, Algorithm 1 does not achieve sublinear regret. We address this by introducing an algorithm, NORA (Algorithm 2), which leverages three crucial ingredients: (1) optimism, (2) off-policy learning, and (3) rare-switching critic updates that target Q^* and accompanying policy resets. Algorithm 2 achieves $\sqrt{dH^4\log|\mathcal{F}|T} + \sqrt{dH^5\log|\mathcal{A}|T}$ regret, requiring only a factor of $dH\log T$ more samples than Algorithm 1 even in the best case for the latter.

Extensions to hybrid RL. Zhou et al. (2023) use both offline and online data to bypass the need to perform strategic exploration in policy optimization. This corresponds to the setting of hybrid RL (Nakamoto et al., 2023; Amortila et al., 2024; Ren et al., 2024; Wagenmaker and Pacchiano, 2023), where Song et al. (2023) show that using both offline and online data allows one to achieve $\sqrt{T}$ regret without optimism. However, the claimed $\sqrt{T}$ regret bound in Zhou et al. (2023) requires on-policy sampling of $O(T)$ samples per timestep that does not contribute to the regret, leading to a sample complexity of $1/\epsilon^6$. Their algorithm cannot achieve sublinear regret in the more common setup where each sample contributes to the regret, while requiring bounded occupancy measure ratios of the optimal policy to *any* policy.

We demonstrate that these issues can be mitigated. Specifically, we extend our optimistic algorithm to leverage both offline and online data, and show that actor-critic methods can benefit from hybrid data and achieve the provable gains in sample efficiency as observed in (Li et al., 2023; Tan and Xu, 2024; Tan et al., 2024). We also provide a non-optimistic provably efficient actor-critic algorithm that only additionally requires $N_{\text{off}} \geq c^*_{\text{off}}(\mathcal{F},\Pi)dH^4/\epsilon^2$ offline samples (with bounded single-policy concentrability) in exchange for omitting optimism. This, along with the result in Theorem 4, shows that hybrid RL therefore allows for sample efficiency gains with optimistic algorithms and computational efficiency gains with non-optimistic algorithms.

Notation. $\mathcal{N}_A(\rho)$ denotes the ρ-covering number of a set A, and $\mathcal{N}_{A,B}(\rho)$ that of $A \cup B$. We use standard asymptotic notation: $f(n) = O(g(n))$ if $f(n)$ grows at most as fast as $g(n)$; $f(n) = o(g(n))$ if it grows strictly slower; $f(n) = \Omega(g(n))$ if it grows at least as fast; and $f(n) = \Theta(g(n))$ if it grows at the same rate.

2. Problem Setting

Markov decision processes. This paper focuses on finite horizon, episodic MDPs, represented by a tuple

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \{\mathbb{P}_h\}_{h=1}^H, \{r_h\}_{h=1}^H),$$

where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, H is the horizon, $r_h : \mathcal{S}\times\mathcal{A} \to [0,1]$ is the reward function at step

[3] Zhong and Zhang (2023) showed that the log-covering number of the policy class increases in the number of actor and critic updates. We sharpen this bound to the number of critic updates.

h and $\mathbb{P}_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel for step h. A policy $\{\pi_h\}_{h=1}^H$ is a set of H functions, where each $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ maps from a state on step h to a probability distribution on actions. Write Π for the class of all policies, and $\pi(s)$ as shorthand for the random variable $a \sim \pi(\cdot|s)$. Given a policy $\{\pi_h\}_{h=1}^H$ and reward function $\{r_h\}_{h=1}^H$, the state value function is defined as

$$V_h^\pi(s) = \mathbb{E}\Big[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'})|s_h = s\Big],$$

where the expectation is taken over the randomness of $a_{h'} \sim \pi_{h'}(s_{h'})$ and $s_{h'+1} \sim \mathbb{P}_h(\cdot|s_{h'}, a_{h'})$ for any $h' \geq h$. The action value, or Q function is defined as

$$Q_h^\pi(s,a) = \mathbb{E}\Big[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'})|s_h = s, a_h = a\Big],$$

where the expectation is taken over the similar randomness of action and state transition, with the only difference that the action randomness is only random as $h' \geq h+1$.

Without loss of generality, write s_1 for the initial state. The optimal policy is $\pi^\star = \operatorname{argmax}_{\pi\in\Pi} V_1^\pi(s_1)$. Correspondingly, we denote $V^\star = V^{\pi^\star}$ and $Q^\star = Q^{\pi^\star}$ as the optimal value and Q-functions. The Bellman operator with respect to the greedy policy and any policy π is given by

$$\mathcal{T}_h Q_{h+1}(s,a) = r_h(s,a) + \mathbb{E}_{s'\sim P_h}\Big[\max_{a'\in\mathcal{A}} Q_{h+1}(s',a')\Big]$$
$$\mathcal{T}_h^\pi Q_{h+1}(s,a) = r_h(s,a) + \mathbb{E}_{s'\sim P_h}\Big[Q_{h+1}(s',\pi_{h+1}(s'))\Big]$$

The optimal Q-function $Q^\star$ is uniquely determined as the solution to the Bellman equation: $Q_h^\star(s,a) = \mathcal{T}_h Q_{h+1}^\star(s,a)$. Our goal is typically to learn an ϵ-optimal policy $\widehat{\pi}$, such that $V_1^*(s_1) - V_1^{\widehat{\pi}}(s_1) \leq \epsilon$, or to obtain sublinear regret over T rounds while playing $(\pi^{(t)})_{t=1}^T$:

$$\mathsf{Reg}(T) = \sum_{t=1}^{T}\Big(V_1^*(s_1) - V_1^{\pi^{(t)}}(s_1)\Big) = o(T).$$

RL with function approximation. Under general function approximation, we approximate Q-functions with a function class $\mathcal{F} = \{\mathcal{F}_h\}_{h\in[H]}$, where each $f_h : \mathcal{S}\times\mathcal{A} \to [0,H]$. The Bellman error with regard to $f \in \mathcal{F}$ is $f_h - \mathcal{T}_h f_{h+1}$, and additionally with regard to $\pi \in \Pi$ as $f_h - \mathcal{T}_h^\pi f_{h+1}$. Additionally, we write $\pi_h^f(a|s) = \mathbb{1}(a' \in \arg\max_{a\in\mathcal{A}} f_h(s,a))$ for the greedy policy that plays the best action under f. We make the following routine assumptions on the richness of $\mathcal{F}$ (Jin et al., 2021; Xie et al., 2022; Rajaraman et al., 2020; Rashidinejad et al., 2023):

Assumption 1 (Realizability)**.** *The function class $\mathcal{F}$ is rich enough such that for all $h \in [H]$, the function class $\mathcal{F}_h$ contains the optimal action value function $Q_h^\star$: $Q_h^* \in \mathcal{F}_h$.*

Assumption 2 (Generalized Completeness)**.** *There exists an auxiliary function class $\mathcal{G} = \mathcal{G}_1 \times ... \times \mathcal{G}_H$, where each $g_h \in \mathcal{G}_h$ satisfies $g_h : \mathcal{S}\times\mathcal{A} \to [0,H]$, that is sufficiently rich such that it contains all Bellman backups of $f \in \mathcal{F}$.*

This auxiliary function class is $(\mathcal{T}^\Pi)^T\mathcal{F} = \{\mathcal{T}^{\pi^{(T)}} \cdot ... \cdot \mathcal{T}^{\pi^{(1)}} f \mid f \in \mathcal{F}, \pi^{(1)}, ..., \pi^{(T)} \in \Pi\}$ for Algorithm 1, and $\mathcal{T}\mathcal{F} = \{\mathcal{T}f \mid f \in \mathcal{F}\}$ for Algorithm 2. The former is far larger than the latter, with exceptions that we highlight in Section 3. We write $\mathcal{N}_\mathcal{F}(\rho)$ for the ρ-covering number of a function class $\mathcal{F}$.[4] To learn $f \in \mathcal{F}$ that approximates the Q-function of a policy π (we say that f targets π), it is common to minimize the temporal difference (TD) error over a dataset $\mathcal{D}$, as an estimate of the Bellman error:

$$\mathcal{L}_h^{(t,\pi)}(f_h, f_{h+1}) = \sum_{(s,a,r,s')\in\mathcal{D}} (f_h(s,a) - r - f_{h+1}(s', \pi_{h+1}(s')))^2. \quad (1)$$

Measures of complexity. The complexity of online learning in the presence of general function approximation is governed by complexity measures such as the Bellman rank (Jiang et al., 2016), which corresponds to the intrinsic dimension in tabular, linear, and linear mixture MDPs. Another is the Bellman eluder dimension (Jin et al., 2021), which subsumes the Bellman rank and additionally characterizes the complexity of kernel, neural, and generalized linear MDPs. We use the squared distributional version:

Definition 1 (Squared Distributional Bellman Eluder dimension (Jin et al., 2021; Zhong et al., 2022; Xiong et al., 2023))**.** *Let $\mathcal{F}$ be a function class. The distributional Bellman Eluder dimension is the largest d such that there exist measures $\{d_h^{(1)}, \ldots, d_h^{(d-1)}\}$, $d_h^{(d)}$, Bellman errors $\{\delta_h^{(1)}, \ldots, \delta_h^{(d-1)}\}$, $\delta_h^{(d)}$, and some $\epsilon' \geq \epsilon$, such that for all $t = 1, ..., d$,*

$$\Big|\mathbb{E}_{d_h^{(t)}}[\delta_h^{(t)}]\Big| > \varepsilon^{(t)} \text{ and } \sqrt{\sum_{i=1}^{t-1}\Big(\mathbb{E}_{d_h^{(i)}}[(\delta_h^{(t)})^2]\Big)} \leq \varepsilon^{(t)}.$$

(Jin et al., 2021) primarily consider two types of distributions: (1) distributions $\mathcal{D}_\mathcal{F}$ induced by greedy policies π^f, and Dirac delta measures over state-action pairs $\mathcal{D}_\Delta$. They suggest an RL problem has low Bellman eluder dimension if either variant is small. Examples include tabular MDPs, where this is the cardinality of the state-action space, and linear MDPs, where this is the corresponding dimension.

The sequential extrapolation coefficient (SEC) of Xie et al. (2022) subsumes the Bellman eluder dimension:

Definition 2 (Sequential Extrapolation Coefficient (SEC))**.**

$$\mathsf{SEC}(\mathcal{F}, \Pi, T) := \max_{h\in[H]} \sup_{f^{(1)},\ldots,f^{(T)}\in\mathcal{F}} \sup_{\pi^{(1)},\ldots,\pi^{(T)}\in\Pi}$$

[4] The ρ-covering number of a class corresponds to the smallest cardinality of a set of points, such that every point in the class is at least ρ-close to some point in that set. See (Wainwright, 2019).

$$\left\{\sum_{t=1}^{T}\frac{\mathbb{E}_{\pi^{(t)}}[f_h^{(t)}-\mathcal{T}_h f_{h+1}^{(t)}]^2}{H^2\vee\sum_{i=1}^{t-1}\mathbb{E}_{\pi^{(i)}}[(f_h^{(t)}-\mathcal{T}_h f_{h+1}^{(t)})^2]}\right\}.$$

The SEC is always bounded by $d\log T$, but there exist MDPs that have a Bellman eluder dimension d on the order of $\sqrt{T}$, but a constant SEC (Xie et al., 2022). We shall use these measures of complexity to characterize the regret. Algorithm 1 scales with the SEC, which is more general and weaker than the Bellman eluder dimension. Algorithm 2, as presented, has a switching cost that scales with the $\mathcal{D}_\Delta$-type Bellman eluder dimension. While this can be weakened to the more general ℓ_2 eluder condition of (Xiong et al., 2023) with nothing more than a change in notation, we present our results in the Bellman eluder dimension framework for familiarity and ease of presentation.

Policy optimization and actor-critic algorithms. Policy optimization approaches optimize directly in the space of policies, enabled by the policy gradient theorem (Sutton and Barto, 2018): $\nabla_\pi V_1^\pi(s_1)=\mathbb{E}_\pi[Q_1^\pi(s,a)\nabla_\pi\log\pi(s,a)]$. This can be done with Monte-Carlo estimates of $Q_1^\pi(s,a)$ (the REINFORCE algorithm), or a learned estimate of $Q_1^\pi(s,a)$ called a critic (actor-critic methods).

However, vanilla policy gradient methods can converge very slowly in the worst case (Li et al., 2021). It is often preferable to use other optimization algorithms, such as a second-order method in natural policy gradient (NPG) (Kakade, 2001), KL-regularized gradient ascent in trust region policy optimization (TRPO) from Schulman et al. (2017), or proximal policy optimization (PPO), which performs a similar, but easier to compute, update. These methods are closely related in the limit of small step-sizes, and are approximate versions of mirror ascent (Schulman et al., 2017; Neu et al., 2017; Rajeswaran et al., 2017).

One instance in which the NPG, TRPO, and PPO updates coincide is with softmax policies (Cai et al., 2024; Cen et al., 2022; Agarwal et al., 2021): $\pi(a|s)=\exp(g(s,a))/\sum_{a\in\mathcal{A}}\exp(g(s,a))$ for some function $g:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$. In this case, the update has a closed form:

$$\pi_h^{(t+1)}(a|s)\propto\pi_h^{(t)}(a|s)\exp(\eta f_h^{(t)}(s,a)),\ f_h^{(t)}\in\mathcal{F}. \quad (2)$$

As mirror ascent, this update is identical to the multiplicative weights or Hedge algorithm. Like (Cai et al., 2024; Zhong and Zhang, 2023; Liu et al., 2023b), we exploit this equivalence to prove our desired regret bounds.

3. Optimistic Actor-Critics – The Easy Case

We now present an optimistic actor-critic algorithm, DOUHUA (Algorithm 1), with provable guarantees under general function approximation. DOUHUA is a natural derivative of the GOLF algorithm of Jin et al. (2021) for actor-critic approaches[5], with only two (very natural) changes. The critic targets $Q^{\pi^{(t)}}$ instead of Q^* while performing optimistic planning at every s,a pair, and we maintain a stochastic policy that is updated with Equation 2. Learning an optimistic critic off-policy achieves sample efficiency by reusing data while exploring efficiently.[6]

Algorithm 1 Double Optimistic Updates for Heavily Updating Actor-critics (DOUHUA)

Input: Function class $\mathcal{F}$.
Initialize: $\mathcal{F}^{(0)}\leftarrow\mathcal{F}$, $\mathcal{D}_h^{(0)}\leftarrow\emptyset$, learning rate $\eta=\Theta(\sqrt{\log|\mathcal{A}|H^{-2}T^{-1}})$, policy $\pi^{(1)}\propto 1$, confidence width $\beta=\Theta(\log(HT\mathcal{N}_{\mathcal{F},(\mathcal{T}^\Pi)^T\mathcal{F}}(1/T)/\delta))$.
for episode $t=1,2,\dots,T$ **do**
  $f_h^{(t)}(s,a)\in\operatorname{argmax}_{f\in\mathcal{F}^{(t-1,\pi^{(t-1)})}}f_h(s,a)\ \forall s,a,h.$
  Play policy $\pi^{(t)}$ for one episode, update dataset $\mathcal{D}_h^{(t)}$.
  Compute confidence set $\mathcal{F}^{(t,\pi^{(t)})}$:

$$\mathcal{F}^{(t,\pi^{(t)})}\leftarrow\Big\{f\in\mathcal{F}:\mathcal{L}_h^{(t,\pi^{(t)})}(f_h,f_{h+1})-\min_{f'_h\in\mathcal{F}_h}\mathcal{L}_h^{(t,\pi^{(t)})}(f'_h,f_{h+1})\le H^2\beta\ \forall h\Big\},$$

$$\mathcal{L}_h^{(t,\pi^{(t)})}(f,f')\leftarrow\sum_{(s,a,r,s')\in\mathcal{D}_h^{(t)}}\Big(f(s,a)-r-f'(s',\pi_{h+1}^{(t)}(s'))\Big)^2.$$

  Update $\pi_h^{(t+1)}(a|s)\propto\pi_h^{(t)}(a|s)\exp(\eta f_h^{(t)}(s,a))$.
end for

Algorithm 1 maintains confidence sets $\mathcal{F}^{(t,\pi^{(t)})}$ that contain all $f\in\mathcal{F}$ where the TD error $\mathcal{L}_h^{(t,\pi^{(t)})}$ with respect to the Bellman operator $\mathcal{T}_h^{\pi^{(t)}}$ is a $H^2\beta$-additive approximation of the minimizer $\min_{f'_h\in\mathcal{F}_h}\mathcal{L}_h^{(t,\pi^{(t)})}(f'_h,f_{h+1})$. Upon the start of each trajectory t, Algorithm 1 maximizes among all functions in the confidence set to play $f_h^{(t)}(s,a)=\sup_{f_h\in\mathcal{F}_h^{(t-1,\pi^{(t-1)})}}f_h(s,a)$ for all h,s,a. This is exactly in line with GOLF, except that the critic targets $Q^{\pi^{(t)}}$ instead of Q^*, and we perform optimistic planning with regard to every state and action like Liu et al. (2023b). We then perform a mirror ascent update $\pi_h^{(t+1)}\propto\pi_h^{(t)}\exp(\eta f_h^{(t)})$ instead of playing the greedy policy $\pi_h^{f^{(t)}}$. This algorithm satisfies the following regret/sample complexity bound:

Theorem 1 (Regret Bound for DOUHUA)**.** *Algorithm 1 achieves the following regret with probability at least* $1-\delta$*:*

$$\mathsf{Reg}(T)=O\left(\sqrt{H^4T\log|\mathcal{A}|}+\sqrt{\beta H^4T\mathsf{SEC}(\mathcal{F},\Pi,T)}\right),$$

where $\beta=\Theta\left(\log\left(HT\mathcal{N}_{\mathcal{F},(\mathcal{T}^\Pi)^T\mathcal{F}}(1/T)/\delta\right)\right)$*. To learn*

[5] Or a completely off-policy version of Liu et al. (2023b).
[6] Similarly to Cai et al. (2024) in linear mixture MDPs.

an ϵ-optimal policy, it therefore requires:

$$N \geq \Omega\left(H^4 \log|\mathcal{A}|/\epsilon^2 + H^4\beta\mathsf{SEC}(\mathcal{F},\Pi,T)/\epsilon^2\right).$$

Proof. We provide a proof sketch here, leaving the full proof to Appendix A. We decompose the regret with Lemma 3:

$$\begin{aligned}
&\mathsf{Reg}(T) \\
&= \underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle f_h^{(t)}(s_h,\cdot), \pi_h^*(\cdot\mid s_h) - \pi_h^{(t)}(\cdot\mid s_h)\right\rangle\right]}_{\text{Tracking error of } \pi^{(t)} \text{ w.r.t. } \pi^*, \text{ bounded by mirror ascent arguments.}} \\
&\quad \underbrace{-\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}\right)(s_h,a_h)\right]}_{\text{Negative Bellman error under } \pi^*, \text{ bounded by optimism.}} \\
&\quad \underbrace{+\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}\right)(s_h,a_h)\right]}_{\text{Bellman error under current policy occupancy, bounded by critic error.}}.
\end{aligned}$$

By a mirror ascent argument in Lemma 4 and our choice of learning rate $\eta = \Theta(\sqrt{\log|\mathcal{A}|H^{-2}T^{-1}})$, the first term is bounded by $\sqrt{H^4T\log|\mathcal{A}|}$.

Lemma 7 establishes optimism: $f_h^{(t)} \geq \mathcal{T}_h^{(t-1)} f_{h+1}^{(t)}$. So we see in Lemma 5 that the second term decomposes into a non-positive term $\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\mathcal{T}_h^{\pi^{(t-1)}} f_{h+1}^{(t)} - f_h^{(t)}\right]$, and $\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} - \mathcal{T}_h^{\pi^{(t-1)}} f_{h+1}^{(t)}\right] \leq \eta H^3 T$.

Bounding the Bellman error under the $\pi^{(t)}$ occupancy measure in Lemma 6 by $\sqrt{\beta H^4 T\mathsf{SEC}(\mathcal{F},\Pi,T)T}$, where $\beta = \Theta\left(\log\left(HT\mathcal{N}_{\mathcal{F},(\mathcal{T}^\Pi)^T\mathcal{F}}(1/T)/\delta\right)\right)$, yields the result. □

A Vacuous Bound. The form of the above regret bound is appealing at first glance. However, the log-covering number of the policy class increases linearly with the number of critic updates by default, as we show below in Lemma 1.[7]

Lemma 1 (Bound on Covering Number of Policy Class, Modified Lemma B.2 from Zhong and Zhang (2023))**.** *Consider the policy class $\Pi^{(T)}$ obtained by performing T updates of the mirror ascent update as in Eq. 2, where the critics $f_h^{(t)} \in \mathcal{F}$ are updated at times $t_1,...,t_K$. Then, the covering number of the policy class at time T is bounded by*

$$\mathcal{N}_{\Pi_h^{(T)}}(\rho/2H) \leq \prod_{k=1}^{K}\mathcal{N}_{\mathcal{F}_h^{(t_k)}}(\rho^2/16\eta K H^2 T).$$

The proof is deferred to Appendix A.5.2. Within Algorithm 1, this implies that $\log\mathcal{N}_{(\mathcal{T}^\Pi)^T\mathcal{F}}(\rho) = O(T\log\mathcal{N}_{\mathcal{F}}(\rho))$ in general, making the bound in Theorem 1 vacuous.

[7] It was previously thought in Zhong and Zhang (2023) that the log-covering number of the policy class increases in the number of actor and critic updates. We sharpen this bound to the number of critic updates in Lemma 1, which may be of independent interest.

A Good Case. However, within certain circumstances, it is possible to have $\log\mathcal{N}_{(\mathcal{T}^\Pi)^T\mathcal{F}}(\rho) = O(\log\mathcal{N}_{\mathcal{F}}(\rho))$. This happens, for instance, when there exists a low-dimensional representation of the sum of clipped Q-functions – such as when the sum of clipped Q-functions is a scaled Q-function:

Definition 3 (Closure Under Truncated Sums)**.** *$\mathcal{F}$ is closed under truncated sums if for any $T \in \mathbb{N}$ and $f^{(1)},...,f^{(T)} \in \mathcal{F}$, $T^{-1}\sum_{t=1}^{T}\min\{\max\{f^{(t)},0\},H\} \in \mathcal{F}$.*

As such, this is an interesting case where the log-covering number of the policy class does not blow up, with downstream implications for the regret of Algorithm 1:

Lemma 2 (Policy Class Growth)**.** *Let $\mathcal{F}$ be a function class that satisfies Definition 3, i.e. it is closed under truncated sums. Then, $\log\mathcal{N}_{(\mathcal{T}^\Pi)^T\mathcal{F}}(\rho) = O(\log\mathcal{N}_{\mathcal{F}}(\rho^2/\eta H^2 T))$.*

Corollary 1 (Regret of DOUHUA, The Good Case)**.** *If $\mathcal{F}$ is closed under truncated sums, then with probability at least $1-\delta$, Algorithm 1 achieves a regret of:*

$$\mathsf{Reg}(T) = O\left(\sqrt{H^4T\log|\mathcal{A}|} + \sqrt{\beta H^4 T\mathsf{SEC}(\mathcal{F},\Pi,T)}\right),$$

where $\beta = \Theta(\log(HT\mathcal{N}_{\mathcal{F}}(H/\log|\mathcal{A}|T^2)/\delta))$. To learn an ϵ-optimal policy, it therefore requires:

$$N \geq \Omega\left(H^4\log|\mathcal{A}|/\epsilon^2 + H^4\beta\mathsf{SEC}(\mathcal{F},\Pi,T)/\epsilon^2\right).$$

We defer the proof of Lemma 2 to Appendix A.5.3. Algorithm 1 then achieves a regret in line with the results of (Efroni et al., 2020; Cai et al., 2024) for tabular and linear mixture MDPs respectively. However, closure under truncated sums is a very strong condition that is not fulfilled by many function classes, although tabular classes fulfill it. It is not fulfilled by linear models due to the clipping operator, requiring Sherman et al. (2024) and Cassel and Rosenberg (2024) to develop bespoke algorithms to get around this in the setting of linear MDPs.[8] With trees, random forests, boosting, and neural networks, without further assumptions, one needs to increase the size of the function class – perhaps even on the same order as the increase in Lemma 1.

This prompts us to explore the possibility of algorithmic modifications to Algorithm 1 in order to achieve the optimal regret rates in the harder, more general case where the log-covering number of the policy class may increase linearly in critic updates. We do so in the next section.

4. Optimistic Actor-Critics – The Hard Case

Given what we have seen in our analysis of Algorithm 1, can we simply modify Algorithm 1 to include rare-switching critic updates? If we can perform only $O(dH\log T)$ critic updates as in (Xiong et al., 2023), perhaps it may be possible to obtain a similar regret bound to Corollary 1.

[8] These avoid truncation by using reward-agnostic exploration and feature shrinkage respectively. The moving target issue is bypassed with rare-switching bonus (but not critic) updates.

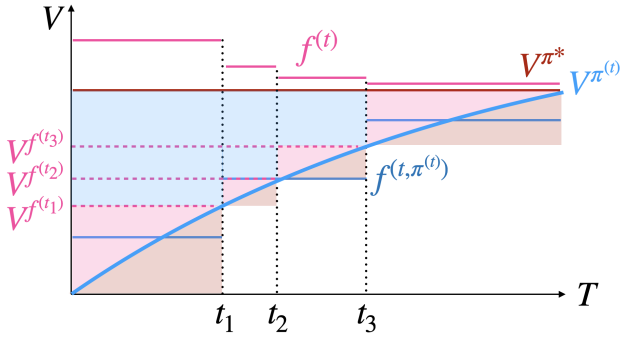

Figure 1. Illustration of tracking error in policy optimization, with a rare-switching critic $f^{(t)}$ that targets π^*. The {blue and pink, pink} area depicts the tracking error of $\{\pi^*$ to $\pi^{(t)}$, $\pi^{f^{(t)}}$ to $\pi^{(t)}\}$. Both incur $\sqrt{T}$ regret. In contrast, $f^{(t,\pi^{(t)})}$ is a rare-switching critic that targets $\pi^{(t)}$, and so is insufficiently optimistic as $\pi^{(t)}$ changes. The blue, pink, and rust area depicts the tracking error of π^* to $\pi^{(t_{\text{last}})}$ from insufficient optimism, which yields linear regret.

4.1. Challenges and algorithm design

However, this is not the case. Intuitively, if the critic targets a rapidly changing $\pi^{(t)}$, the Bellman error with regard to the policy $\pi^{(t_{\text{last}})}$ at the last update will not be close to the Bellman error with regard to $\pi^{(t)}$. Although the former is what the critic $f^{(t_{\text{last}})} = f^{(t)}$ targets, we evaluate the latter when considering a switch at time t. Therefore, the critic updates far more often when we target $Q^{\pi^{(t)}}$ rather than Q^* – as it tries to hit a moving target.

Furthermore, this results in insufficient optimism. In Lemma 7, we can only guarantee optimism with respect to the Bellman operator at the last critic update time t_{last}, $f_h^{(t)} \geq \mathcal{T}_h^{\pi^{(t_{\text{last}})}} f_{h+1}^{(t)}$. But we require $f_h^{(t)} \geq \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}$. Attempting to work with this form of limited optimism results in the tracking error relative to π^* becoming

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\left\langle f_h^{(t)}(s',\cdot), \pi_h^*(\cdot|s') - \pi_h^{(t_{\text{last}})}(\cdot|s')\right\rangle\right],$$

incurring linear regret. This is the shaded area in Figure 1.

A way forward. Having the critic target Q^* instead of $Q^{\pi^{(t)}}$ provides a solution. This ensures sufficient optimism, as we show in Lemma 14 that $f_h^{(t)} \geq \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}$. Furthermore, we do not need to control $\log \mathcal{N}_{(\mathcal{T}^\Pi)^T\mathcal{F}}(\rho)$, as $\max_a Q_{h+1}(s,a)$ is a contraction and it suffices to control $\log \mathcal{N}_{\mathcal{T}\mathcal{F}}(\rho)$, which is often approximately $\log \mathcal{N}_{\mathcal{F}}(\rho)$.

However, this introduces an additional term we need to control – the deviation of the current policy $\pi^{(t)}$ from the greedy policy $\pi^{f^{(t)}}$, depicted as the pink area in Figure 1. This term is difficult to control, as $\pi^{f^{(t)}}$ changes with every critic update, and the actor requires sufficient time to catch up to the critic updates. To address this issue, we introduce policy resets at every critic update, allowing us to bound this with the standard mirror descent regret bound as in Lemma 12. The total tracking error scales with the number of critic updates, which is then resolved with rare-switching critic updates in line with Xiong et al. (2023). Lastly, we increase the learning rate to reduce the regret incurred by the policy resets. This makes more aggressive policy updates to catch up with the sudden, rare, and large critic updates.

Summary. Algorithm 2 combines *optimism* for strategic exploration and *off-policy learning* for sample efficiency. While rare-switching critic updates are a-priori appealing, they are unstable when the critic targets $\pi^{(t)}$, due to limited optimism and the challenge of tracking a moving policy (as we describe below, and further show in Appendix F). These are resolved by *targeting* π^* and re-introducing *rare-switching critic updates* respectively. However, this introduces additional error, as the agent has to effectively unlearn the policy after each rare critic update. We therefore introduce *policy resets* to the uniform policy after each critic update – incurring minimal cost due to their rarity (as there are only $O(dH\log T)$ updates). A *more aggressive learning rate* (by a factor of $\sqrt{dH\log T}$) mitigates some of the additional regret incurred, and can be seen as making more aggressive updates to make up for lost ground from policy resets. See Appendix B for more details.

4.2. Regret bound for NORA

To control the switching cost with the framework of Xiong et al. (2023), we concern ourselves with function classes of low D_Δ-type Bellman Eluder dimension (Jin et al., 2021):

Assumption 3 (Bounded D_Δ-type Bellman Eluder Dimension)**.** *Let $\mathcal{D}_\Delta := \{\mathcal{D}_{\Delta,h}\}_{h\in[H]}$, where $\mathcal{D}_{\Delta,h} = \{\delta_{(s,a)}(\cdot) \mid s\in\mathcal{S}, a\in\mathcal{A}\}$. That is, we only consider distributions that are Dirac deltas on a single state-action pair. We assume that $d := d_{BE}(\mathcal{F}, D_\Delta, 1/\sqrt{T}) < \infty$.*

This can be weakened to the more general ℓ_2 eluder condition of (Xiong et al., 2023) with nothing more than a change in notation. However, we present our results in the Bellman eluder dimension framework for familiarity and ease of presentation. We now show the following:

Theorem 2 (Regret Bound for NORA)**.** *Algorithm 2 achieves the following regret with probability at least $1-\delta$:*

$$\mathsf{Reg}(T) = O\left(\sqrt{dH^5T\log T\log|\mathcal{A}|} + dH^3\log T + \sqrt{\beta H^4 T \mathsf{SEC}(\mathcal{F},\Pi,T)}\right).$$

where $\beta = \Theta(\log(HT^2\mathcal{N}_{\mathcal{F},\mathcal{T}\mathcal{F}}(1/T)/\delta))$. This implies a sample complexity (ignoring lower order terms) of

$$N \geq \Omega\left(dH^5\log T\log|\mathcal{A}|/\epsilon^2 + H^4\beta\mathsf{SEC}(\mathcal{F},\Pi,T)/\epsilon^2\right).$$

Algorithm 2 No-regret Optimistic Rare-switching Actor-critic (NORA)

Input: Function class $\mathcal{F}$.
Initialize: $\mathcal{F}^{(0)} \leftarrow \mathcal{F}$, $\mathcal{D}_h^{(0)} \leftarrow \emptyset, \forall h \in [H]$, $\eta = \Theta(\sqrt{d \log T \log |\mathcal{A}| H^{-1} T^{-1}})$, $\pi^{(1)} \propto 1$, confidence width $\beta = \Theta(\log(HT^2 \mathcal{N}_{\mathcal{F},\mathcal{TF}}(1/T)/\delta))$.
for episode $t = 1, 2, \ldots, T$ **do**
  Set $f_h^{(t)}(s,a) \in \operatorname{argmax}_{f \in \mathcal{F}^{(t_{\text{last}})}} f_h(s,a) \; \forall s, a, h$.
  Play policy $\pi^{(t)}$ for one episode, obtain trajectory, update dataset $\mathcal{D}_h^{(t)}$.
  if $\mathcal{L}_h^{(t)}(f_h^{(t)}, f_{h+1}^{(t)}) \geq \min_{f_h' \in \mathcal{F}_h} \mathcal{L}_h^{(t)}(f_h', f_{h+1}^{(t)}) + 5H^2\beta$ for some h **then**
    Compute confidence set $\mathcal{F}^{(t)}$:

$$\mathcal{F}^{(t)} \leftarrow \left\{ f \in \mathcal{F} : \mathcal{L}_h^{(t)}(f_h, f_{h+1}) - \min_{f_h' \in \mathcal{F}_h} \mathcal{L}_h^{(t)}(f_h', f_{h+1}) \leq H^2\beta \; \forall h \right\},$$

$$\mathcal{L}_h^{(t)}(f, f') \leftarrow \sum_{(s,a,r,s') \in \mathcal{D}_h^{(t)}} \left(f(s,a) - r - \max_{a' \in \mathcal{A}} f'(s', a') \right)^2.$$

    Reset policy $\pi^{(t)} \propto 1$.
    Set $t_{\text{last}} \leftarrow t$, $N_{\text{updates}}^{(t)} \leftarrow N_{\text{updates}}^{(t-1)} + 1$.
  else
    Set $N_{\text{updates}}^{(t)} \leftarrow N_{\text{updates}}^{(t-1)}$, $\mathcal{F}^{(t)} \leftarrow \mathcal{F}^{(t-1)}$.
  end if
  Update $\pi_h^{(t+1)}(a|s) \propto \pi_h^{(t)}(a|s) \exp(\eta f_h^{(t)}(s,a))$.
end for

Proof. We provide a proof sketch here, and defer the full proof to Appendix B. By Lemma 8, we have a slightly different regret decomposition than usual:

$$\begin{aligned}
&\mathsf{Reg}(T) \\
&= \underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\left\langle f_h^{(t)}(s_h,\cdot), \pi_h^*(\cdot \mid s_h) - \pi_h^{(t)}(\cdot \mid s_h)\right\rangle\right]}_{\text{Tracking error of } \pi^{(t)} \text{ w.r.t. } \pi^*, \text{ bounded by mirror ascent arguments.}} \\
&\quad \underbrace{- \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}\right)(s_h,a_h)\right]}_{\text{Negative Bellman error under } \pi^*, \text{ bounded by optimism.}} \\
&\quad \underbrace{+ \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)} - \mathcal{T}_h f_{h+1}^{(t)}\right)(s_h,a_h)\right]}_{\text{Bellman error under current policy occupancy, bounded by critic error.}} \\
&\quad \underbrace{+ \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^{(t)}}\left[\left\langle f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^{f^{(t)}}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')\right\rangle\right]}_{\text{Tracking error of } \pi^{(t)} \text{ w.r.t. } \pi^{f^{(t)}}, \text{ bounded by mirror ascent and policy resets.}}.
\end{aligned}$$

We take care of the second and third terms first. Targeting Q^* yields sufficient optimism to assert that $f_h^{(t)} \geq \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}$ in Lemma 14, and so we argue that second term for the negative Bellman error under π^* is non-positive in Lemma 10. The third term is bounded by $\sqrt{\beta H^4 T \mathsf{SEC}(\mathcal{F}, \Pi, T)}$ via a standard argument from Xie et al. (2022) in Lemma 11.

Instead of directly bounding the first and fourth terms right away, consider the critic switch times $t_1, ..., t_K$. Then,

$$\sum_{t=t_k+1}^{t_{k+1}} \sum_{h=1}^{H} \mathbb{E}_{\pi^{(t)}}\left[\left\langle f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^{f^{(t)}}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')\right\rangle\right]$$

is bounded by $O(\eta H^3(t_{k+1} - t_k) + H \log|\mathcal{A}|/\eta + H^2)$ in Lemma 12. We exploit the fact that we reset the policy after the episode is collected at every t_k in doing so. We then do the same for the other term in Lemma 9.

To conclude, we bound the number of critic updates in Lemma 13 by $dH \log T$ with the techniques of Xiong et al. (2023). Summing over all $t_1, ..., t_K$ lets us bound the first and fourth terms by $\sqrt{dH^5 \log T \log|\mathcal{A}|} + dH^3 \log T$ by our choice of learning rate. We modify the learning rate $\eta = \Theta(\sqrt{d \log T \log|\mathcal{A}| H^{-1} T^{-1}})$ to mitigate the switch cost, shaving off a factor of $dH \log T$ in the final sample complexity. Putting everything together yields the result. □

Quality of the regret bound. Even compared to the "good case" for Algorithm 1 in Corollary 1, Algorithm 2 requires only $dH \log T$ more samples to learn an ϵ-optimal policy. This is exactly the same as the switch cost, in line with what Cassel and Rosenberg (2024) observe for linear MDPs.

The Bellman eluder dimension can scale poorly with T (e.g., $d_{\text{BE}}(\mathcal{F}, D_\Delta, 1/\sqrt{T}) = \Omega(\sqrt{T})$ in rare cases), and so we primarily rely on the SEC of Xie et al. (2022), which is always at most $O(d \log T)$ – though our result still depends on d due to the switch cost. Still, function classes with low Bellman eluder dimension are ubiquitous, and so Algorithm 2 achieves $\sqrt{T}$ regret on a large class of problems including linear and kernel MDPs (Jin et al., 2021). We can use the weaker ℓ_2 eluder condition of (Xiong et al., 2023) with only a change in notation, but we present our results in the language of the Bellman eluder dimension for clarity.

Comparison with other work. The only other method claiming $\sqrt{T}$ regret with policy optimization and general function approximation is Zhou et al. (2023). However, they allow themselves to collect $\Omega(T)$ samples *at every timestep t without incurring any additional regret*. This incurs a sample complexity of $1/\epsilon^6$, while Algorithm 2 enjoys a $1/\epsilon^2$ gurarantee in comparison. Their regret bound is not sublinear in the more common setup where this sampling contributes to the regret. Liu et al. (2023b) achieves $1/\epsilon^3$ complexity with batched critic updates but do not quantify the potential growth of the policy class.

Extension to other policy updates. We are able to accommodate other policy optimization updates other than the multiplicative weights update, if they satisfy the following:

Corollary 2. *If there exists some policy optimization oracle* $\pi^{(t+1)} \leftarrow PO(\pi^{(t)}, f^{(t)}, \eta)$*, and some* $\mathsf{OPT}^*, \mathsf{OPT}^f$ *such that* $\sum_{h,t} \mathbb{E}_{\pi^*}[\langle f_h^{(t)}(s,\cdot), \pi_h^*(\cdot|s) - \pi_h^{(t)}(\cdot|s)\rangle] \le \mathsf{OPT}^*$, $\sum_{h,t} \mathbb{E}_{\pi^{(t)}}[\langle f_h^{(t)}(s,\cdot), \pi_h^{f^{(t)}}(\cdot|s) - \pi_h^{(t)}(\cdot|s)\rangle] \le \mathsf{OPT}^f$, *then Algorithm 2 with this update obtains a regret of* $\mathsf{Reg}(T) = O\left(\sqrt{\beta H^4 T \mathsf{SEC}(\mathcal{F}, \Pi, T)} + \mathsf{OPT}^* + \mathsf{OPT}^f\right)$.

4.3. Further comments on the design of NORA

Targeting Q^*. We are not the only ones who consider a critic that targets Q^* instead of $Q^{\pi^{(t)}}$. For example, Crites and Barto (1994) propose an actor-critic algorithm that mimics Q-learning via a critic that targets Q^*. Similarly, popular methods like DDPG (Lillicrap et al., 2019) and TD3 (Fujimoto et al., 2018) alternately update a deterministic policy approximately maximizing its own Q-function estimate (thus indirectly targeting Q^*), combined with stochastic exploration to track the evolving policy.

Similarity to deep deterministic policy gradients (DDPG). Algorithm 2 shares similarities with DDPG (Lillicrap et al., 2019), which alternates updates between a deterministic policy $g_h^{(t)}(s) \approx \arg\max_a f_h^{(t)}(s,a)$ and a critic minimizing the TD error, using slowly updated targets reminiscent of our rare-switching updates. DDPG explores via Gaussian perturbations around its deterministic policy, indirectly targeting Q^*. This suggests that DDPG and its successor TD3 (Fujimoto et al., 2018) may be useful backbones for adapting algorithmic insights garnered from the design and analysis of Algorithm 2 for practical RL.[9]

5. Extension to Hybrid RL

Here, we demonstrate two benefits of hybrid offline-online data in actor-critic algorithms. With optimism, these algorithms achieve improved sample efficiency. Alternatively, offline data enables one to omit optimism, simplifying implementation and improving computational efficiency.

5.1. Computational efficiency through hybrid RL

Song et al. (2023); Amortila et al. (2024); Zhou et al. (2023) show hybrid RL can avoid optimism altogether given offline data with sufficient coverage. Following this idea, we provide non-optimistic variants of Algorithms 1 and 2 in Algorithms 3 and 4. These achieve $\sqrt{T}$ regret without optimizing over a confidence set, detailed in Appendix C.

[9]This provides insight on why DDPG/TD3 are more sample-efficient than on-policy PPO. Algorithm 2 needs $1/\epsilon^2$ samples, while on-policy sampling needs at least $1/\epsilon^4$ (Liu et al., 2023b)

When $c^*_{\text{off}}(\mathcal{F}, \Pi)$ is the single-policy concentrability coefficient defined in Definition 4, these enjoy the following guarantees (with proof deferred to Appendix D.1):

Theorem 3. *Algorithms 3 and 4 achieve an additional regret of* $O(\sqrt{\beta^\pi H^4 c^*_{\text{off}}(\mathcal{F}, \Pi) T^2 / N_{\text{off}}})$ *and* $O(\sqrt{\beta^* H^4 c^*_{\text{off}}(\mathcal{F}) T^2 / N_{\text{off}}})$ *respectively over Algorithms 1 and 2, where* $\beta^\pi = \Theta\left(\log\left(HT\mathcal{N}_{\mathcal{F},(\mathcal{T}^\Pi)^T\mathcal{F}}(1/T)/\delta\right)\right)$, $\beta^* = \Theta(\log\left(HT^2\mathcal{N}_{\mathcal{F},\mathcal{T}\mathcal{F}}(1/T)/\delta\right))$.

In exchange for omitting optimism, we incur an additional $\sqrt{\beta H^4 c^*_{\text{off}}(\mathcal{F}) T^2 / N_{\text{off}}}$ error term, which amounts to $\sqrt{T}$ regret as long as $N_{\text{off}} = \Omega(T)$. This shows that the provable efficiency achieved without optimism by hybrid FQI-type methods (Song et al., 2023) also extends to actor-critic methods – resolving the issue within Zhou et al. (2023) of needing to collect $\Theta(T)$ samples at every timestep t. We achieve a sample complexity of $1/\epsilon^2$, in contrast to the $1/\epsilon^6$ sample complexity of Zhou et al. (2023).

5.2. Sample efficiency gains with hybrid RL

The extension, found in Algorithm 5, appends the offline data to the online data and minimizes the TD error over the combined dataset when constructing the confidence sets. This yields the following guarantee:

Theorem 4 (Hybrid RL Regret Bound for NORA)**.** *Let* $\mathcal{X}_{\text{off}}, \mathcal{X}_{\text{on}}$ *be an arbitrary partition over* $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times [H]$. *Algorithm 5 satisfies with probability at least* $1 - \delta$*:*

$$\mathsf{Reg}(N_{\text{on}}) = \mathcal{O}\left(\inf_{\mathcal{X}_{\text{on}}, \mathcal{X}_{\text{off}}} \left(\sqrt{\beta H^4 c_{\text{off}}(\mathcal{F}\mathbb{1}_{\mathcal{X}_{\text{off}}}) N_{\text{on}}^2 / N_{\text{off}}} + \sqrt{\beta H^4 N_{\text{on}} \mathsf{SEC}(\mathcal{F}\mathbb{1}_{\mathcal{X}_{\text{on}}}, \Pi, T)} + \sqrt{dH^5 N_{\text{on}} \log|\mathcal{A}|}\right)\right),$$

where $\beta = C \log(HN\mathcal{N}_{\mathcal{F},\mathcal{T}\mathcal{F}}(1/N)/\delta)$ *for some constant* C, $N = N_{\text{off}} + N_{\text{on}}$, $\mathbb{1}_{\mathcal{X}_{\text{off}}}, \mathbb{1}_{\mathcal{X}_{\text{on}}}$ *are indicator variables for whether* $s, a \in \mathcal{X}_{\text{off}}$ *or* $\mathcal{X}_{\text{on}}$*, and* $c_{\text{off}}(\mathcal{F}\mathbb{1}_{\mathcal{X}_{\text{off}}})$ *is the partial all-policy concentrability coefficient (Tan and Xu, 2024).*

We defer the proof to Appendix D.3. Broadly, the critic error splits into offline and online terms, bounded by offline coverage and online exploration respectively. Algorithm 5 is completely unaware of the partition, but the regret bound optimizes over all partitions of the state-action space. With sufficient high-quality offline data, the regret is approximately $\sqrt{\beta H^4 N_{\text{on}} \mathsf{SEC}(\mathcal{F}\mathbb{1}_{\mathcal{X}_{\text{on}}}, \Pi, T)} + \sqrt{dH^5 N_{\text{on}} \log|\mathcal{A}|}$ for some small $\mathcal{X}_{\text{on}}$, yielding improvements over Theorem 2. This shows that actor-critic methods can benefit from hybrid data, achieving the provable gains in sample efficiency compared to offline-only and online-only learning observed by Li et al. (2023); Tan et al. (2024).

6. Numerical experiments

We provide two numerical experiments to empirically verify our findings, with details deferred to Appendix H.

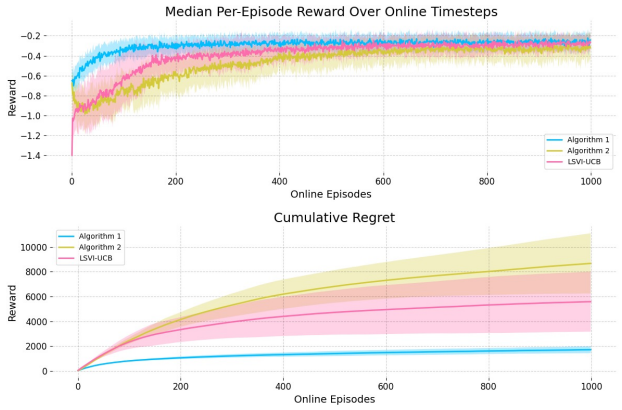

Figure 2. Per-episode reward and cumulative regret of Algorithms 1 and 2, compared to a rare-switching version of LSVI-UCB (Jin et al., 2019) on a linear MDP tetris task. Algorithm 1 outperforms LSVI-UCB, and Alg. 2 catches up after some time, though all achieve $\sqrt{T}$ regret. Results averaged over 30 trials.

Optimism in linear MDPs. The first experiment examines Algorithms 1 and 2 in a linear MDP setting, in order to validate if they indeed achieve $\sqrt{T}$ regret in practice. Accordingly, we implement optimism with LSVI-UCB bonuses (Jin et al., 2019) instead of global optimism as in GOLF. We compare our algorithms to a rare-switching[10] version of LSVI-UCB on a linear MDP tetris task (Tan et al., 2024; Tan and Xu, 2024). The condition required for Algorithm 1 to work holds here, as we do not clip the Q-function estimates. Figure 2 shows that Algorithm 1 (surprisingly) performs better than LSVI-UCB, and empirically illustrates that Algorithm 2 also achieves $\sqrt{T}$ regret even though it performs slightly worse than LSVI-UCB.

Deep hybrid RL. We compare variants of Algorithms 3 and 4, that we call Algorithms 1H and 2H respectively, to Cal-QL (Nakamoto et al., 2023). The differences are as follows. Algorithms 1H and 2H employ offline pretraining of both the actor and the critic, and the actor employs the soft actor-critic (SAC) update from (Haarnoja et al., 2018). Additionally, Algorithm 2H omits the policy resets. Like Cal-QL, and unlike Algorithm 1H, Algorithm 2H takes a maximum over 10 randomly sampled actions from the policy to enhance exploration – approximately targeting π^*.

Algorithm 2H outperforms Cal-QL, which in turn slightly outperforms Algorithm 1H. These results suggest that Algorithms 1H and 2H remain highly competitive, and may perform just as well as the state of the art in hybrid RL (Cal-QL) – even without the use of pessimism. Our results support our theoretical findings that hybrid RL allows for computationally efficient actor-critic algorithms.

[10]Implemented with the doubling determinant method used in (He et al., 2023). This incurs at most a $dH \log T$ switching cost.

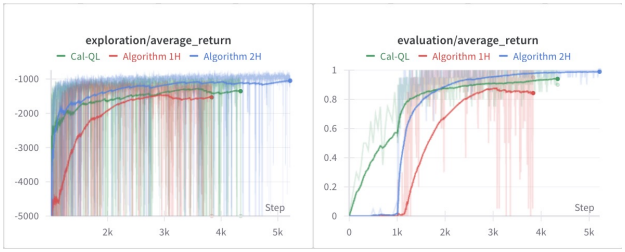

Figure 3. Comparison of Cal-QL (Nakamoto et al., 2023), Alg. 1H, and Alg. 2H on the `antmaze-medium-diverse-v2` task. Alg. 2H outperforms both Cal-QL and Alg. 1H, suggesting that hybrid RL enables efficient exploration without pessimism. Evaluation plots show offline pretraining in the first 1000 steps. All plots are exponentially smoothed.

7. Conclusions and Future Work

We resolve an open problem in the online RL literature by designing an actor-critic method in Algorithm 2 with general function approximation that achieves $1/\epsilon^2$ (and matching $\sqrt{T}$ regret) without making any reachability or coverage assumptions. This was achieved through several key ingredients in the algorithm design. By (1) performing optimistic exploration, (2) learning a critic off-policy without throwing away any data, (3) having the critic target Q^* instead of $Q^{\pi^{(t)}}$ to ensure sufficient optimism and enable rare-switching critic updates, and (4) performing policy resets to control the deviation of the current policy from the greedy policy, we achieve the desired result in Theorem 2.

We resolve another open problem in hybrid RL with a non-optimistic, provably efficient actor-critic algorithm requiring $N_{\text{off}} \geq c^*_{\text{off}}(\mathcal{F}, \Pi) dH^4/\epsilon^2$ offline samples. Together with Theorem 4, this shows hybrid RL enables sample efficiency via optimism and computational efficiency without it.

While Algorithm 2 is generally computationally inefficient, the insights gained are promising for empirical applications. Optimism, crucial for strategic exploration, can be implemented via linear bonuses (Agarwal et al., 2020), count-based methods (Martin et al., 2017), randomized value functions (Osband et al., 2019), or random latent exploration (Mahankali et al., 2024). An off-policy critic can leverage DDQN (van Hasselt et al., 2015), updating only on high TD error. Optimistic TD3 (Fujimoto et al., 2018) with rare-switching updates and resets is particularly promising. Further numerical experiments would be valuable.

One might incorporate rarely-updated bonus functions (Agarwal et al., 2022; Zhao et al., 2023), updating the base critic each episode. This may achieve optimism without targeting Q^* if sums of Q-functions approximate scaled Q-functions. Extending He et al. (2023) to study minimax-optimal policy optimization in linear MDPs as Wu et al. (2022) did for tabular MDPs, is another valuable direction.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

The authors are supported in part by the NSF grants CCF-2106778, CCF-2418156 and CAREER award DMS-2143215.

References

Abbasi-Yadkori, Y., Bartlett, P., Bhatia, K., Lazic, N., Szepesvari, C., and Weisz, G. (2019). POLITEX: Regret bounds for policy iteration using expert prediction. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3692–3702. PMLR.

Agarwal, A., Henaff, M., Kakade, S., and Sun, W. (2020). Pc-pg: Policy cover directed exploration for provable policy gradient learning.

Agarwal, A., Jin, Y., and Zhang, T. (2022). Vo*q*l: Towards optimal regret in model-free rl with nonlinear function approximation.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2021). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76.

Amortila, P., Foster, D. J., Jiang, N., Sekhari, A., and Xie, T. (2024). Harnessing density ratios for online reinforcement learning.

Bhandari, J. and Russo, D. (2022). Global optimality guarantees for policy gradient methods.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. (2024). Provably efficient exploration in policy optimization.

Cassel, A. and Rosenberg, A. (2024). Warm-up free policy optimization: Improved regret in linear markov decision processes.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. (2022). Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578.

Chen, Z., Li, C. J., Yuan, A., Gu, Q., and Jordan, M. I. (2022). A general framework for sample-efficient function approximation in reinforcement learning.

Crites, R. and Barto, A. (1994). An actor/critic algorithm that is equivalent to q-learning. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems*, volume 7. MIT Press.

Efroni, Y., Shani, L., Rosenberg, A., and Mannor, S. (2020). Optimistic policy optimization with bandit feedback.

Fujimoto, S., van Hoof, H., and Meger, D. (2018). Addressing function approximation error in actor-critic methods.

Gaur, M., Bedi, A. S., Wang, D., and Aggarwal, V. (2024). Closing the gap: Achieving global convergence (last iterate) of actor-critic under markovian sampling with neural network parametrization.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor.

He, J., Zhao, H., Zhou, D., and Gu, Q. (2023). Nearly minimax optimal reinforcement learning for linear markov decision processes.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. (2016). Contextual decision processes with low bellman rank are pac-learnable.

Jin, C., Liu, Q., and Miryoosefi, S. (2021). Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2019). Provably efficient reinforcement learning with linear function approximation.

Kakade, S. M. (2001). A natural policy gradient. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Kumar, N., Agrawal, P., Ramponi, G., Levy, K. Y., and Mannor, S. (2024). Improved sample complexity for global convergence of actor-critic algorithms.

Lan, G. (2022). Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2021). Softmax policy gradient methods can take exponential time to converge. In *Conference on Learning Theory*, pages 3107–3110. PMLR.

Li, G., Zhan, W., Lee, J. D., Chi, Y., and Chen, Y. (2023). Reward-agnostic fine-tuning: Provable statistical benefits of hybrid reinforcement learning. *arXiv preprint arXiv:2305.10282*.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2019). Continuous control with deep reinforcement learning.

Liu, B., Cai, Q., Yang, Z., and Wang, Z. (2023a). Neural proximal/trust region policy optimization attains globally optimal policy.

Liu, Q., Weisz, G., György, A., Jin, C., and Szepesvári, C. (2023b). Optimistic natural policy gradient: a simple efficient policy optimization framework for online rl.

Mahankali, S., Hong, Z.-W., Sekhari, A., Rakhlin, A., and Agrawal, P. (2024). Random latent exploration for deep reinforcement learning.

Martin, J., Sasikumar, S. N., Everitt, T., and Hutter, M. (2017). Count-based exploration in feature space for reinforcement learning.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. (2022). On the global convergence rates of softmax policy gradient methods.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning.

Munos, R. and Szepesvári, C. (2008). Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(27):815–857.

Nakamoto, M., Zhai, Y., Singh, A., Mark, M. S., Ma, Y., Finn, C., Kumar, A., and Levine, S. (2023). Cal-ql: Calibrated offline rl pre-training for efficient online fine-tuning.

Neu, G., Jonsson, A., and Gómez, V. (2017). A unified view of entropy-regularized markov decision processes.

Osband, I., Roy, B. V., Russo, D., and Wen, Z. (2019). Deep exploration via randomized value functions.

Rajaraman, N., Yang, L. F., Jiao, J., and Ramachandran, K. (2020). Toward the fundamental limits of imitation learning.

Rajeswaran, A., Kumar, V., Gupta, A., Vezzani, G., Schulman, J., Todorov, E., and Levine, S. (2017). Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *arXiv preprint arXiv:1709.10087*.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. (2023). Bridging offline reinforcement learning and imitation learning: A tale of pessimism.

Ren, J., Swamy, G., Wu, Z. S., Bagnell, J. A., and Choudhury, S. (2024). Hybrid inverse reinforcement learning.

Schulman, J., Levine, S., Moritz, P., Jordan, M. I., and Abbeel, P. (2017). Trust region policy optimization.

Sherman, U., Cohen, A., Koren, T., and Mansour, Y. (2024). Rate-optimal policy optimization for linear markov decision processes.

Song, Y., Zhou, Y., Sekhari, A., Bagnell, J. A., Krishnamurthy, A., and Sun, W. (2023). Hybrid rl: Using both offline and online data can make rl efficient.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. The MIT Press, second edition.

Tan, K., Fan, W., and Wei, Y. (2024). Hybrid reinforcement learning breaks sample size barriers in linear mdps. In *Advances in Neural Information Processing Systems*.

Tan, K. and Xu, Z. (2024). A natural extension to online algorithms for hybrid RL with limited coverage. *Reinforcement Learning Journal*, 1.

van Hasselt, H., Guez, A., and Silver, D. (2015). Deep reinforcement learning with double q-learning.

Wagenmaker, A. and Pacchiano, A. (2023). Leveraging offline data in online reinforcement learning.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Wu, T., Yang, Y., Zhong, H., Wang, L., Du, S. S., and Jiao, J. (2022). Nearly optimal policy optimization with stable at any time guarantee.

Xiao, L. (2022). On the convergence rates of policy gradient methods.

Xie, T., Foster, D. J., Bai, Y., Jiang, N., and Kakade, S. M. (2022). The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157*.

Xiong, N., Wang, Z., and Yang, Z. (2023). A general framework for sequential decision-making under adaptivity constraints.

Yuan, R., Du, S. S., Gower, R. M., Lazaric, A., and Xiao, L. (2023). Linear convergence of natural policy gradient methods with log-linear policies.

Zanette, A., Cheng, C.-A., and Agarwal, A. (2021). Cautiously optimistic policy optimization and exploration with linear function approximation.

Zhan, W., Huang, B., Huang, A., Jiang, N., and Lee, J. D. (2022). Offline reinforcement learning with realizability and single-policy concentrability.

Zhao, H., He, J., and Gu, Q. (2023). A nearly optimal and low-switching algorithm for reinforcement learning with general function approximation.

Zhong, H., Xiong, W., Zheng, S., Wang, L., Wang, Z., Yang, Z., and Zhang, T. (2022). Gec: A unified framework for interactive decision making in mdp, pomdp, and beyond. *arXiv preprint arXiv:2211.01962*.

Zhong, H. and Zhang, T. (2023). A theoretical analysis of optimistic proximal policy optimization in linear markov decision processes.

Zhou, Y., Sekhari, A., Song, Y., and Sun, W. (2023). Offline data enhanced on-policy policy gradient with provable guarantees.

Contents

A. Proofs for Theorem 1

We prove a regret bound for Algorithm 1 here. This algorithm achieves sublinear regret only when the covering number of the function class does not increase linearly with the number of critic updates.

A.1. Regret decomposition

We first work with the following regret decomposition below. This is the same regret decomposition as that of Cai et al. (2024) and Zhong and Zhang (2023), and we provide the proof for completeness.

Lemma 3 (Regret Decomposition For Q^π-Targeting Actor-Critics)**.** *The regret at time T yields the following decomposition*

$$\begin{aligned}\mathrm{Reg}(T) = &\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle f_h^{(t)}(s_h,\cdot), \pi_h^*(\cdot\mid s_h) - \pi_h^{(t)}(\cdot\mid s_h)\right\rangle\right] - \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^*} f_{h+1}^{(t)}\right)(s_h,a_h)\right] \\ &+ \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}\right)(s_h,a_h)\right].\end{aligned}$$

Proof. By adding and subtracting $f_1^{(t)}(s_1^{(t)}, \pi_1^{(t)}(s_1^{(t)}))$, we obtain

$$\begin{aligned}\mathrm{Reg}(T) &= \sum_{t=1}^{T}\left(V_1^*(s_1^{(t)}) - V_1^{\pi^{(t)}}(s_1^{(t)})\right) \\ &= \sum_{t=1}^{T}\left(V_1^*(s_1^{(t)}) - f_1^{(t)}(s_1^{(t)}, \pi_1^{(t)}(s_1^{(t)}))\right) + \sum_{t=1}^{T}\left(f_1^{(t)}(s_1^{(t)}, \pi_1^{(t)}(s_1^{(t)})) - V_1^{\pi^{(t)}}(s_1^{(t)})\right). \end{aligned} \tag{3}$$

We can now apply the value difference lemma/generalized policy difference lemma in Lemma 23 (Cai et al., 2024; Efroni et al., 2020) with $f^{(t)}$ as the Q-function, $\pi' = \pi^*$, and $\pi = \pi^{(t)}$ to find that

$$\begin{aligned}&\sum_{t=1}^{T}\left(V_1^*(s_1^{(t)}) - f_1^{(t)}(s_1^{(t)}, \pi_1^{(t)}(s_1^{(t)}))\right) \\ &= \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle f_h^{(t)}(s_h,\cdot), \pi_h^*(\cdot\mid s_h) - \pi_h^{(t)}(\cdot\mid s_h)\right\rangle\right] - \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^*} f_{h+1}^{(t)}\right)(s_h,a_h)\right].\end{aligned} \tag{4}$$

Another application of Lemma 23 with $\pi = \pi' = \pi^{(t)}$ and with $f^{(t)}$ as the Q-function yields

$$\begin{aligned}&\sum_{t=1}^{T}\left(f_1^{(t)}(s_1^{(t)}, \pi_1^{(t)}(s_1^{(t)})) - V_1^{\pi^{(t)}}(s_1^{(t)})\right) \\ &= \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{(t)}}\left[\left\langle f_h^{(t)}(s_h,\cdot), \pi_h^{(t)}(\cdot\mid s_h) - \pi_h^{(t)}(\cdot\mid s_h)\right\rangle\right] + \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}\right)(s_h,a_h)\right] \\ &= \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}\right)(s_h,a_h)\right].\end{aligned} \tag{5}$$

Plugging (4) and (5) into (3) concludes the proof. □

A.2. Bounding the tracking error

The first term is bounded by $\sqrt{H^4 T \log|\mathcal{A}|}$, as we see in the following lemma. This lemma is the standard mirror descent regret bound, and is a similar argument to lemmas found in Liu et al. (2023b) and Cai et al. (2024).

Lemma 4 (Mirror Descent Tracking Error for Algorithm 1)**.** *Let $\pi^{(1)} \propto 1$ and consider updating policies with respect to a set of Q-function estimates $f^{(1)}, ..., f^{(T)}$ by*

$$\pi_{h+1}^{(t+1)}(\cdot|s') = \frac{\pi_{h+1}^{(t)}(\cdot|s')\exp(\eta f_{h+1}^{(t)}(s',\cdot))}{\sum_{a\in\mathcal{A}}\pi_{h+1}^{(t)}(a|s')\exp(\eta f_{h+1}^{(t)}(s',a))} = Z^{-1}\pi_{h+1}^{(t)}(\cdot|s')\exp(\eta f_{h+1}^{(t)}(s',\cdot)). \tag{6}$$

The tracking error with respect to the optimal policy is then bounded by:

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle f_h^{(t)}\left(s_h,\cdot\right), \pi_h^*\left(\cdot \mid s_h\right) - \pi_h^{(t)}\left(\cdot \mid s_h\right)\right\rangle\right] \leq \eta H^3 T/2 + \frac{H\log|\mathcal{A}|}{\eta}.$$

Proof. Rearranging (6) yields

$$\eta f_{h+1}^{(t)}(s',\cdot) = \log Z + \log\pi_{h+1}^{(t+1)}(\cdot|s') - \log\pi_{h+1}^{(t)}(\cdot|s'),$$

where $\log Z$ is

$$\log Z = \log\left(\sum_{a\in\mathcal{A}}\pi_{h+1}^{(t)}(a|s')\exp(\eta f_{h+1}^{(t)}(s',a))\right).$$

We can now bound, noting that $\sum_{a\in\mathcal{A}}\left(\pi_{h+1}^*(\cdot|s') - \pi_{h+1}^{(t+1)}(\cdot|s')\right) = 0$, that

$$\begin{aligned}
&\left\langle \eta f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^*(\cdot|s') - \pi_{h+1}^{(t+1)}(\cdot|s')\right\rangle \\
&= \left\langle \log Z + \log\pi_{h+1}^{(t+1)}(\cdot|s') - \log\pi_{h+1}^{(t)}(\cdot|s'), \pi_{h+1}^*(\cdot|s') - \pi_{h+1}^{(t+1)}(\cdot|s')\right\rangle \\
&= \left\langle \log\pi_{h+1}^{(t+1)}(\cdot|s') - \log\pi_{h+1}^{(t)}(\cdot|s'), \pi_{h+1}^*(\cdot|s') - \pi_{h+1}^{(t+1)}(\cdot|s')\right\rangle \\
&= \mathrm{KL}\left(\pi_{h+1}^*(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^*(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^{(t+1)}(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right),
\end{aligned} \tag{7}$$

where the last line follows directly from Lemma 24 by setting $\pi = \pi_{h+1}^{(t+1)}(\cdot|s')$, $\pi_1 = \pi_{h+1}^\star(\cdot|s')$ and $\pi_2 = \pi_{h+1}^{(t)}(\cdot|s')$. As a result, we can bound the desired inner product as follows,

$$\begin{aligned}
&\left\langle \eta f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^*(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')\right\rangle \\
&= \left\langle \eta f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^*(\cdot|s') - \pi_{h+1}^{(t+1)}(\cdot|s')\right\rangle - \left\langle \eta f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^{(t)}(\cdot|s') - \pi_{h+1}^{(t+1)}(\cdot|s')\right\rangle \\
&\leq \mathrm{KL}\left(\pi_{h+1}^*(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^*(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^{(t+1)}(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right) \\
&\quad + \eta H||\pi_{h+1}^{(t+1)}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')||_1,
\end{aligned} \tag{8}$$

where the last line follows directly from (7). Summing up the inner product bounded in (8), we derive that

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^*(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')\right\rangle\right] \\
&= \frac{1}{\eta}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle \eta f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^*(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')\right\rangle\right] \\
&\leq \frac{1}{\eta}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\Big[\mathrm{KL}\left(\pi_{h+1}^*(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^*(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right) \\
&\qquad\qquad -\mathrm{KL}\left(\pi_{h+1}^{(t+1)}(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right) + \eta H||\pi_{h+1}^{(t+1)}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')||_1\Big].
\end{aligned} \tag{9}$$

Now we apply Pinsker's inequality and note that

$$\mathrm{KL}(\pi_{h+1}^{(t+1)}(\cdot|s')||\pi_{h+1}^{(t)}(\cdot|s')) \geq \|\pi_{h+1}^{(t+1)}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')\|_1^2/2, \tag{10}$$

and plugging (10) into (9) yields that

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^*(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')\right\rangle\right] \\
&\leq \frac{1}{\eta}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\Big[\mathrm{KL}\left(\pi_{h+1}^*(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^*(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right) \\
&\qquad\qquad -||\pi_{h+1}^{(t+1)}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')||_1^2/2 + \eta H||\pi_{h+1}^{(t+1)}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')||_1\Big] \\
&\leq \frac{1}{\eta}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\mathrm{KL}\left(\pi_{h+1}^*(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^*(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right) + \eta^2 H^2/2\right],
\end{aligned} \tag{11}$$

where we use the fact that $\max_{x\in\mathbb{R}}\left\{-x^2/2 + \eta H x\right\} = \eta^2 H^2/2$ in the last line. Continuing to simplify this expression yields

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^*(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')\right\rangle\right] \\
&\leq \sum_{t=1}^{T}\sum_{h=1}^{H}\left(\eta H^2/2 + \mathbb{E}_{\pi^*}\left[\frac{\mathrm{KL}\left(\pi_{h+1}^*(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^*(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right)}{\eta}\right]\right) \\
&\leq \eta H^3 T/2 + \sum_{h=1}^{H}\frac{\mathrm{KL}\left(\pi_{h+1}^*(\cdot|s') \,||\, \pi_{h+1}^{(1)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^*(\cdot|s') \,||\, \pi_{h+1}^{(T)}(\cdot|s')\right)}{\eta} \\
&\leq \eta H^3 T/2 + \frac{H\log|\mathcal{A}|}{\eta},
\end{aligned} \tag{12}$$

where the last inequality follows from the fact that the KL-divergence is non-negative as well as noting that $\pi^{(1)}$ is the uniform policy, so that the KL divergence can be bounded as

$$\begin{aligned}
\mathrm{KL}\left(\pi_{h+1}^*(\cdot|s') \,||\, \pi_{h+1}^{(1)}(\cdot|s')\right) &= \sum_{a\in\mathcal{A}}\pi_{h+1}^*(a|s')\log\pi_{h+1}^*(a|s') - \sum_{a\in\mathcal{A}}\pi_{h+1}^*(a|s')\log\pi_{h+1}^{(1)}(a|s') \\
&\leq -\sum_{a\in\mathcal{A}}\pi_{h+1}^*(a|s')\log\pi_{h+1}^{(1)}(a|s') \\
&= \log|\mathcal{A}|.
\end{aligned} \tag{13}$$

□

A.3. Asserting optimism

Lemma 5 (Negative Bellman Error, Algorithm 1)**.** *Within Algorithm 1, it holds that*

$$-\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[f_h^{(t)} - \mathcal{T}_h^{\pi^*}f_{h+1}^{(t)}\right] \leq \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle f_{h+1}^{(t)}(s_{h+1},\cdot), \pi_{h+1}^*(\cdot|s_{h+1}) - \pi_{h+1}^{(t)}(\cdot|s_{h+1})\right\rangle\right] + \eta H^3 T.$$

Proof. First, we decompose the negative bellman error as follows

$$-\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[f_h^{(t)} - \mathcal{T}_h^{\pi^*}f_{h+1}^{(t)}\right] = \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\mathcal{T}_h^{\pi^{(t)}}f_{h+1}^{(t)} - f_h^{(t)}\right] + \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\mathcal{T}_h^{\pi^*}f_{h+1}^{(t)} - \mathcal{T}_h^{\pi^{(t)}}f_{h+1}^{(t)}\right]$$

$$= \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\mathcal{T}_h^{\pi^{(t-1)}} f_{h+1}^{(t)} - f_h^{(t)}\right] + \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} - \mathcal{T}_h^{\pi^{(t-1)}} f_{h+1}^{(t)}\right] + \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\mathcal{T}_h^{\pi^*} f_{h+1}^{(t)} - \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}\right]. \tag{14}$$

Here, using the result from Lemma 7 that $f_h^{(t)} \geq \mathcal{T}_h^{(t-1)} f_{h+1}^{(t)}$, we note that the first term is less than 0. We employ Lemma 25 with $t_1 = 1$ and $t_2 = T$ to bound the second term by $\eta H^3 T$. For the third term, we note that

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\mathcal{T}_h^{\pi^*} f_{h+1}^{(t)} - \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}\right] \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[r_h(s_h,a_h) + f_{h+1}^{(t)}\left(s_{h+1}, \pi_{h+1}^*(s_{h+1})\right) - r_h(s_h,a_h) - f_{h+1}^{(t)}\left(s_{h+1}, \pi_{h+1}^{(t)}(s_{h+1})\right)\right] \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[f_{h+1}^{(t)}\left(s_{h+1}, \pi_{h+1}^*(s_{h+1})\right) - f_{h+1}^{(t)}\left(s_{h+1}, \pi_{h+1}^{(t)}(s_{h+1})\right)\right] \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\left\langle f_{h+1}^{(t)}(s_{h+1},\cdot), \pi_{h+1}^*(\cdot|s_{h+1}) - \pi_{h+1}^{(t)}(\cdot|s_{h+1})\right\rangle\right].
\end{aligned} \tag{15}$$

Replacing the last term in (14) with (15) concludes the proof. □

A.4. Bounding the Bellman error under the learned policies

Lemma 6 (Bellman Error Under Policy Occupancy, Algorithm 1)**.** *Within Algorithm 1, it holds that*

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}\right)(s_h,a_h)\right] \leq \sqrt{\beta H^4 T \mathsf{SEC}(\mathcal{F},\Pi,T)} + \eta H^3 T$$

with probability at least $1-\delta$*, where* $\beta = \Theta\left(\log\left(TG\mathcal{N}_{\mathcal{F},(\mathcal{T}^{\Pi})^T\mathcal{F}}(1/T)/\delta\right)\right)$.

Proof. We first decompose the Bellman error

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}\right)(s_h,a_h)\right] \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t-1)}} f_{h+1}^{(t)}\right)(s_h,a_h)\right] + \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^{(t)}}\left[\left(\mathcal{T}_h^{\pi^{(t-1)}} f_{h+1}^{(t)} - \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}\right)(s_h,a_h)\right] \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t-1)}} f_{h+1}^{(t)}\right)(s_h,a_h)\right] + \eta H^3 T,
\end{aligned} \tag{16}$$

where the last line holds by Lemma 25. We will further bound the first term by applying Cauchy-Schwarz inequality, and we will see that

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t-1)}} f_{h+1}^{(t)}\right)(s_h,a_h)\right] \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t-1)}} f_{h+1}^{(t)}\right)(s_h,a_h)\right] \left(\frac{H^2 \vee \sum_{i=1}^{t-1} \mathbb{E}_{\pi^{(i)}}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t-1)}} f_{h+1}^{(t)}\right)^2\right]}{H^2 \vee \sum_{i=1}^{t-1} \mathbb{E}_{\pi^{(i)}}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t-1)}} f_{h+1}^{(t)}\right)^2\right]}\right)^{1/2}
\end{aligned}$$

$$\leq \sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H} \frac{\mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t-1)}} f_{h+1}^{(t)}\right)\right]^2}{H^2 \vee \sum_{i=1}^{t-1} \mathbb{E}_{\pi^{(i)}}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t-1)}} f_{h+1}^{(t)}\right)^2\right]}} \sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H} H^2 \vee \sum_{i=1}^{t-1} \mathbb{E}_{\pi^{(i)}}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t-1)}} f_{h+1}^{(t)}\right)^2\right]}, \quad (17)$$

where the last step follows from Cauchy-Schwarz inequality. Within the last inequality, the first term can be bounded by H times the SEC of Xie et al. (2022), by the very definition of the SEC in Definition 2. The second term is bounded by Lemma 14, where $\beta = \Theta\left(\log\left(TG\mathcal{N}_{\mathcal{F},(\mathcal{T}^{\Pi})^T\mathcal{F}}(1/T)/\delta\right)\right)$. Putting these bounds into (17), we obtain that

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t-1)}} f_{h+1}^{(t)}\right)(s_h, a_h)\right] \leq \sqrt{H\mathsf{SEC}(\mathcal{F}, \Pi, T)}\sqrt{\beta H^3 T} \leq \sqrt{\beta H^4 T \mathsf{SEC}(\mathcal{F}, \Pi, T)}. \quad (18)$$

Plugging (18) into (16), we finally obtain that

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}\right)(s_h, a_h)\right] \leq \sqrt{\beta H^4 T \mathsf{SEC}(\mathcal{F}, \Pi, T)} + \eta H^3 T. \quad (19)$$

□

A.5. Auxiliary lemmas

A.5.1. Showing optimism for critics targeting $Q^{\pi^{(t)}}$

We prove the following lemma in more generality than is needed for Algorithm 1, accomodating critic updates that are rarer than in every episode, with the aim to use the more general result in future sections. This can be thought of as an analogue of Lemma 15 in Xie et al. (2022), which is also Lemmas 39 and 40 in Jin et al. (2021).

Lemma 7 (Optimism and in-sample error control for critics targeting $Q^{\pi^{(t)}}$)**.** *Let t_{last} be the time of the last critic update before episode t. Consider a critic targeting $Q^{\pi^{(t)}}$ as in Algorithm 1. With probability at least $1-\delta$, for all $t \in [T]$, we have that for all $h = 1, \ldots, H$*

$$\textit{(i)}\ \mathcal{T}_h^{\pi^{(t_{last})}} f_{h+1}^{(t)} \in \mathcal{F}_h^{(t)}, \textit{ and so } f_h^{(t)} \geq \mathcal{T}_h^{\pi^{(t_{last})}} f_{h+1}^{(t)},$$
$$\textit{(ii)}\ \sum_{i=1}^{t-1} \mathbb{E}_{d^{\pi^{(i)}}}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t-1)}} f_{h+1}^{(t)}\right)^2\right] \leq O\left(H^2\beta\right),$$

by choosing $\beta = c_1(\log[HT\mathcal{N}_{\mathcal{F},(\mathcal{T}^{\Pi})^T\mathcal{F}}(\rho)/\delta])$ for some constant c_1.

Proof. We note that it does not hold that $Q^* \in \mathcal{F}^{(t)}$, as our Bellman operator is given by the operator $\mathcal{T}_h^{\pi^{(t)}}$ under policy $\pi^{(t)}$ and not the greedy policy under $f^{(t)}$. Furthermore, as we do not throw away samples not in the current batch as Liu et al. (2023b) do, we do not enjoy the same conditional independence of dataset and next-step value functions. We therefore take a different approach, of modifying the analysis of Xiong et al. (2023) to the policy gradient setting in order to do so.

By Lemma 17 applied to policy $\pi^{(t)}$, for any $h \in [H]$ and $t \in [T]$, we have with probability $1-\delta$ that

$$0 \leq \mathcal{L}_h^{(t,\pi^{(t)})}(\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}, f_{h+1}^{(t)}) - \min_{f_h' \in \mathcal{F}_h} \mathcal{L}_h^{(t)}(f_h', f_{h+1}^{(t)}) \leq H^2\beta. \quad (20)$$

We construct the confidence sets only at timesteps t_k where switches occur:

$$\mathcal{F}_h^{(t_{\text{last}},\pi^{(t_{\text{last}})})} := \left\{ f \in \mathcal{F} : \mathcal{L}_h^{(t_{\text{last}},\pi^{(t_{\text{last}})})}(f_h, f_{h+1}) - \min_{f_h' \in \mathcal{F}_h} \mathcal{L}_h^{(t_{\text{last}},\pi^{(t_{\text{last}})})}(f_h', f_{h+1}) \leq H^2\beta \ \ \forall h \in [H] \right\}. \quad (21)$$

It must then follow that $\mathcal{T}_h^{\pi^{(t_k)}} f_{h+1} \in \mathcal{F}_h^{(t_k)}$ for any $f \in \mathcal{F}$ and t_k. Now, we want to establish that

$$\mathcal{T}_h^{\pi^{(t_{\text{last}})}} f_{h+1}^{(t)} \leq \sup_{f_h \in \mathcal{F}_h^{(t_{\text{last}})}} f_h(s,a) = f_h^{(t)}. \quad (22)$$

Further recall that we have defined

$$\mathcal{F}_h^{(t_{\text{last}})} := \left\{ f \in \mathcal{F} : \mathcal{L}_h^{(t_{\text{last}})}(f_h, f_{h+1}) - \min_{f'_h \in \mathcal{F}_h} \mathcal{L}_h^{(t_{\text{last}})}(f'_h, f_{h+1}) \leq H^2\beta \ \ \forall h \in [H] \right\}.$$

So we can apply Lemma 17 on t_{last}, $f_{h+1}^{(t)}$, and $\pi^{(t)}$ to find that

$$0 \leq \mathcal{L}_h^{(t_{\text{last}})}(\mathcal{T}_h^{\pi^{(t_{\text{last}})}} f_{h+1}^{(t)}, f_{h+1}^{(t)}) - \min_{f'_h \in \mathcal{F}_h} \mathcal{L}_h^{(t_{\text{last}})}(f'_h, f_{h+1}^{(t)}) \leq H^2\beta. \tag{23}$$

It then holds that

$$\mathcal{T}_h^{\pi^{(t_{\text{last}})}} f_{h+1}^{(t)} \in \mathcal{F}_h^{t_{\text{last}}}, \text{ and so } \mathcal{T}_h^{\pi^{(t_{\text{last}})}} f_{h+1}^{(t)} \leq \sup_{f_h \in \mathcal{F}_h^{(t_{\text{last}})}} f_h(s,a) = f_h^{(t)}.$$

The second result can now be shown by Lemma 18 using a similar argument to the proof of Theorem 1 in Xiong et al. (2023), which in turn takes inspiration from the proofs of Lemmas 39 and 40 in Jin et al. (2021). We elaborate accordingly.

Consider two cases, one where we perform an update at episode $t-1$ and one where we do not. If we perform an update at episode $t-1$, then by the construction of $\mathcal{F}^{(t-1)}$ it must hold that

$$\mathcal{L}_h^{(t-1,\pi^{(t-1)})}\left(f_h^{(t)}, f_{h+1}^{(t)}\right) - \min_{f'_h \in \mathcal{F}_h} \mathcal{L}_h^{(t-1,\pi^{(t-1)})}\left(f'_h, f_{h+1}^{(t)}\right) \leq H^2\beta,$$

and by Lemma 17 it must also hold that

$$0 \leq \mathcal{L}_h^{(t-1,\pi^{(t-1)})}\left(\mathcal{T}_h^{\pi^{(t-1)}} f_h^{(t)}, f_{h+1}^{(t)}\right) - \min_{f'_h \in \mathcal{F}_h} \mathcal{L}_h^{(t-1,\pi^{(t-1)})}\left(f'_h, f_{h+1}^{(t)}\right) \leq H^2\beta.$$

One can then see that

$$\mathcal{L}_h^{(t-1,\pi^{(t-1)})}\left(f_h^{(t)}, f_{h+1}^{(t)}\right) - \mathcal{L}_h^{(t-1,\pi^{(t-1)})}\left(\mathcal{T}_h^{\pi^{(t-1)}} f_h^{(t)}, f_{h+1}^{(t)}\right) \leq 6H^2\beta. \tag{24}$$

The same holds for the other case where we do not perform an update at episode $t-1$. Observe that because we did not perform an update,

$$\mathcal{L}_h^{(t-1,\pi^{(t-1)})}\left(f_h^{(t-1)}, f_{h+1}^{(t-1)}\right) - \min_{f'_h \in \mathcal{F}_h} \mathcal{L}_h^{(t-1,\pi^{(t-1)})}\left(f'_h, f_{h+1}^{(t-1)}\right) \leq 5H^2\beta.$$

From Lemma 17, it also holds that

$$\mathcal{L}_h^{(t-1,\pi^{(t-1)})}\left(\mathcal{T}_h^{\pi^{(t-1)}} f_h^{(t-1)}, f_{h+1}^{(t-1)}\right) - \min_{f'_h \in \mathcal{F}_h} \mathcal{L}_h^{(t-1,\pi^{(t-1)})}\left(f'_h, f_{h+1}^{(t-1)}\right) \leq H^2\beta.$$

Putting the above two statements together and using that $f^{(t)} = f^{(t-1)}$ again yields

$$\begin{aligned}
&\mathcal{L}_h^{(t-1,\pi^{(t-1)})}\left(f_h^{(t)}, f_{h+1}^{(t)}\right) - \mathcal{L}_h^{(t-1,\pi^{(t-1)})}\left(\mathcal{T}_h^{\pi^{(t-1)}} f_h^{(t)}, f_{h+1}^{(t)}\right) \\
&= \mathcal{L}_h^{(t-1,\pi^{(t-1)})}\left(f_h^{(t-1)}, f_{h+1}^{(t-1)}\right) - \mathcal{L}_h^{(t-1,\pi^{(t-1)})}\left(\mathcal{T}_h^{\pi^{(t-1)}} f_h^{(t-1)}, f_{h+1}^{(t-1)}\right) \\
&\leq 6H^2\beta.
\end{aligned} \tag{25}$$

An application of Lemma 18 to both cases then yields in either case that

$$\sum_{i=1}^{t-1} \left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t-1)}} f_{h+1}^{(t)}\right)^2 (s_h^{(t)}, a_h^{(t)}) \leq 7H^2\beta, \tag{26}$$

and also that

$$\sum_{i=1}^{t-1} \mathbb{E}_{d_h^{(i)}} \left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t-1)}} f_{h+1}^{(t)}\right)^2\right] \leq 7H^2\beta. \tag{27}$$

□

A.5.2. Bound on covering number of value function class (Proof of Lemma 1)

The proof is similar to that of Lemma B.2 in Zhong and Zhang (2023), but we strengthen the result to show that the covering number increases only in the number of critic updates, not policy updates. As in Zhong and Zhang (2023), let $\mathcal{F}_{\rho^2/16\eta KH^2T,h}^{(t_k)}$ be a minimal $\rho^2/16\eta KH^2T$-net of $\mathcal{F}_h^{(t_k)}$ for all k. So for any $\pi \propto \exp(\eta \sum_{i=1}^K (t_i - t_{i-1}) f_h^{(t_i)})$ with $f_h^{(t_i)} \in \mathcal{F}_h^{(t_i)}$, there exists some $\widehat{f}_h^{(t_i)} \in \mathcal{F}_{\rho^2/16\eta KH^2T,h}^{(t_k)}$ so

$$\sup_{s,a} |f_h^{(t_i)}(s,a) - \widehat{f}_h^{(t_i)}(s,a)| \leq \frac{\rho^2}{16\eta KH^2T} \text{ for all } i \in [K].$$

It then holds that

$$\begin{aligned}\sup_{s,a} \left| \eta \sum_{i=1}^K (t_i - t_{i-1}) f_h^{(t_i)}(s,a) - \eta \sum_{i=1}^K (t_i - t_{i-1}) \widehat{f}_h^{(t_i)}(s,a) \right| &\leq \eta \sum_{i=1}^K (t_i - t_{i-1}) \sup_{s,a} |f_h^{(t_i)}(s,a) - \widehat{f}_h^{(t_i)}(s,a)| \\ &\leq \eta T \sum_{i=1}^K \sup_{s,a} |f_h^{(t_i)}(s,a) - \widehat{f}_h^{(t_i)}(s,a)| \leq \frac{\rho^2}{16H^2}. \end{aligned} \tag{28}$$

Now we invoke Lemma B.3 of Zhong and Zhang (2023), provided as Lemma 21 for completeness, to show that

$$\sup_s ||\pi(\cdot|s) - \pi'(\cdot|s)||_1 \leq 2\sqrt{\sup_{s,a} \left| \eta \sum_{i=1}^K (t_i - t_{i-1}) f_h^{(t_i)}(s,a) - \eta \sum_{i=1}^K (t_i - t_{i-1}) \widehat{f}_h^{(t_i)}(s,a) \right|} \leq \frac{\rho}{2H}. \tag{29}$$

It then holds that

$$\mathcal{N}_{\Pi_h^{(T)}}(\rho/2H) \leq \prod_{i=1}^K \mathcal{N}_{\mathcal{F}_h^{(t_i)}}(\rho^2/16\eta KH^2T). \tag{30}$$

The bound for $\mathcal{N}_{\mathcal{A}}(\rho/2H)$ follows from discretizing the action space $\mathcal{A}$ via a covering number argument, and observing that the covering number of an $\mathcal{N}_{\mathcal{A}}(\rho)$-dimensional probability distribution is on the order of $\mathcal{N}_{\mathcal{A}}(\rho)$.

A.5.3. Closure under truncated sums limits policy class growth (Proof of Lemma 2)

Let $\mathcal{F}_{\rho^2/16\eta TH^2}$ be a minimal $\rho^2/16\eta TH^2$-net of $\mathcal{F}$. Then, for all $f_h \in \mathcal{F}$, there exists some $\widehat{f}_h \in \mathcal{F}_{\rho^2/16\eta TH^2}$ such that

$$\sup_{s,a} |f_h(s,a) - \widehat{f}_h(s,a)| \leq \frac{\rho^2}{16\eta TH^2}.$$

As $\mathcal{F}$ is closed under truncated sums, it then holds that there exists some $f' \in \mathcal{F}$ such that

$$\begin{aligned}\sup_{s,a} \left| \eta \sum_{t=1}^T f_h^{(t)}(s,a) - \eta \sum_{t=1}^T \widehat{f}_h(s,a) \right| &= \sup_{s,a} \left| \eta \sum_{t=1}^T \min\{\max\{f_h^{(t)}(s,a), 0\}, H\} - \eta \sum_{t=1}^T \widehat{f}_h(s,a) \right| \\ &= \sup_{s,a} \left| \eta T f'_h(s,a) - \eta T \widehat{f}_h(s,a) \right| \leq \frac{\rho^2}{16H^2}. \end{aligned} \tag{31}$$

The result then follows from Lemma 20 and Lemma 21. That is, Lemma 21 shows that

$$\sup_s ||\pi(\cdot|s) - \pi'(\cdot|s)||_1 \leq 2\sqrt{\sup_{s,a} \left| \eta \sum_{t=1}^T f_h^{(t)}(s,a) - \eta \sum_{t=1}^T \widehat{f}_h^{(t)}(s,a) \right|} \leq \frac{\rho}{2H}, \tag{32}$$

it then holds that

$$\mathcal{N}_{\Pi_h^{(T)}}(\rho/2H) \leq \mathcal{N}_{\mathcal{F}}(\rho^2/16\eta TH^2), \tag{33}$$

and the result directly follows from Lemma 20.

B. Proofs for Theorem 2

Recall the motivation for this solution:

1. **Optimism** allows one to perform strategic exploration, addressing the issue of exploration vs. exploitation, and allowing us to avoid making reachability or coverage assumptions.
2. **Off-policy learning** avoids throwing away any samples, ensuring that no samples are wasted.
3. **Do rare-switching critic updates work?** A-priori, introducing rare-switching critic updates as in Xiong et al. (2023) offers an appealing solution to the covering number issue. However, the Bellman operator with respect to $\pi^{(t)}$ has a very limited form of optimism. Furthermore, it is difficult to control the number of rare-switching updates in the context of general function approximation, where we make an update when the Bellman error with respect to the current policy is large, as the current policy keeps changing and so we track a moving target.
4. **Letting the critic target Q^* and not Q^π** ensures sufficient optimism, as the Bellman operator is now the same at every iteration. Further, as the critic targets Q^*, we do not need to control $\log \mathcal{N}_{(\mathcal{T}^\Pi)^T \mathcal{F}}(\rho)$, as the Bellman operator for the greedy policy is a contraction.
5. **Rare-switching critic updates now work.** However, this introduces an additional term, where we need to bound the deviation of the current policy from its target, the greedy policy with regard to the current critic. Re-introducing rare-switching critic updates resolves this, as now we allow the actor sufficient time to catch up to the critic updates. Controlling the number of critic updates is not an issue when the critic targets Q^*, as the Bellman operator is now the same at every iteration. This is reminiscent of the delayed target Q-function trick common in deep RL (Lillicrap et al., 2019; Fujimoto et al., 2018).
6. **Policy resets.** However, going through the mirror descent proof to control the additional error results in an additional term bounded by $\sum_{k=1}^{N_{\text{updates}}} \log(1/\pi^{(t_k)})$. This term can be controlled by resetting the policy to the uniform policy upon every critic update. As critic updates are rare, on the order of $dH \log T$, the additional error incurred is very small. This trick was adopted independently by Cassel and Rosenberg (2024) for the same reason.
7. **Increased learning rate.** We increase the learning rate accordingly by a factor of $\sqrt{dH \log T}$, exactly the square root of the number of critic updates/policy resets. This is done to mitigate the increase in regret incurred by the policy resets by a factor of $\sqrt{dH \log T}$. This can be seen as making more aggressive updates to make up for the lost ground due to policy resets when the critic makes a rare but large update.

Given the above, we are now in a position to continue our analysis below.

B.1. Regret decomposition

We employ the following regret decomposition below. This is a slightly different regret decomposition than that of Cai et al. (2024) and Zhong and Zhang (2023), as our critic targets Q^*.

Lemma 8 (Regret Decomposition For Q^*-Targeting Actor-Critics)**.**

$$\begin{aligned}\mathrm{Reg}(T) = &\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\left\langle f_h^{(t)}(s_h,\cdot), \pi_h^*(\cdot \mid s_h) - \pi_h^{(t)}(\cdot \mid s_h)\right\rangle\right] - \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^*} f_{h+1}^{(t)}\right)(s_h,a_h)\right] \\ &+ \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)} - \mathcal{T}_h f_{h+1}^{(t)}\right)(s_h,a_h)\right] + \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^{(t)}}\left[\left\langle f_{h+1}^{(t)}(s_{h+1},\cdot), \pi_{h+1}^{f^{(t)}}(\cdot|s_{h+1}) - \pi_{h+1}^{(t)}(\cdot|s_{h+1})\right\rangle\right].\end{aligned}$$

Proof. By adding and subtracting $f_1^{(t)}(s_1^{(t)}, \pi_1^{(t)}(s_1^{(t)}))$, we obtain

$$\begin{aligned}\mathrm{Reg}(T) &= \sum_{t=1}^{T}\left(V_1^*(s_1^{(t)}) - V_1^{\pi^{(t)}}(s_1^{(t)})\right) \\ &= \sum_{t=1}^{T}\left(V_1^*(s_1^{(t)}) - f_1^{(t)}(s_1^{(t)}, \pi_1^{(t)}(s_1^{(t)}))\right) + \sum_{t=1}^{T}\left(f_1^{(t)}(s_1^{(t)}, \pi_1^{(t)}(s_1^{(t)})) - V_1^{\pi^{(t)}}(s_1^{(t)})\right). \qquad (34)\end{aligned}$$

To further decompose the two terms in (34), we apply the value difference lemma/generalized policy difference lemma in Lemma 23 (Cai et al., 2024; Efroni et al., 2020) with $f^{(t)}$ as the Q-function, $\pi' = \pi^*$, and $\pi = \pi^{(t)}$ to find that the first term can written as

$$\begin{aligned}
&\sum_{t=1}^{T}\left(V_1^*(s_1^{(t)}) - f_1^{(t)}(s_1^{(t)}, \pi_1^{(t)}(s_1^{(t)}))\right) \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle f_h^{(t)}(s_h,\cdot), \pi_h^*(\cdot \mid s_h) - \pi_h^{(t)}(\cdot \mid s_h)\right\rangle\right] - \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^*} f_{h+1}^{(t)}\right)(s_h, a_h)\right]. \qquad (35)
\end{aligned}$$

Another application of Lemma 23 with $\pi = \pi' = \pi^{(t)}$ and with $f^{(t)}$ as the Q-function yields

$$\begin{aligned}
&\sum_{t=1}^{T}\left(f_1^{(t)}(s_1^{(t)}, \pi_1^{(t)}(s_1^{(t)})) - V_1^{\pi^{(t)}}(s_1^{(t)})\right) \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{(t)}}\left[\left\langle f_h^{(t)}(s_h,\cdot), \pi_h^{(t)}(\cdot \mid s_h) - \pi_h^{(t)}(\cdot \mid s_h)\right\rangle\right] + \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}\right)(s_h, a_h)\right] \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}\right)(s_h, a_h)\right]. \qquad (36)
\end{aligned}$$

We now need to relate the Bellman operator with respect to the current policy $\mathcal{T}_h^{\pi^{(t)}}$ to the Bellman operator $\mathcal{T}_h$. Observe that

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}\right)(s_h, a_h)\right] \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)} - \mathcal{T}_h f_{h+1}^{(t)}\right)(s_h, a_h)\right] + \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{(t)}}\left[\left(\mathcal{T}_h f_{h+1}^{(t)} - \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}\right)(s_h, a_h)\right]. \qquad (37)
\end{aligned}$$

We further decompose the second term in the following way

$$\begin{aligned}
&\mathbb{E}_{\pi^{(t)}}\left[\left(\mathcal{T}_h f_{h+1}^{(t)} - \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}\right)(s_h, a_h)\right] \\
&= \mathbb{E}_{\pi^{(t)}}\left[r_h(s_h, a_h) + f_{h+1}^{(t)}\left(s_{h+1}, \pi_{h+1}^{f^{(t)}}(s_{h+1})\right) - r_h(s_h, a_h) - f_{h+1}^{(t)}\left(s_{h+1}, \pi_{h+1}^{(t)}(s_{h+1})\right)\right] \\
&= \mathbb{E}_{\pi^{(t)}}\left[f_{h+1}^{(t)}\left(s_{h+1}, \pi_{h+1}^{f^{(t)}}(s_{h+1})\right) - f_{h+1}^{(t)}\left(s_{h+1}, \pi_{h+1}^{(t)}(s_{h+1})\right)\right] \\
&= \mathbb{E}_{\pi^{(t)}}\left[\left\langle f_{h+1}^{(t)}(s_{h+1},\cdot), \pi_{h+1}^{f^{(t)}}(\cdot|s_{h+1}) - \pi_{h+1}^{(t)}(\cdot|s_{h+1})\right\rangle\right]. \qquad (38)
\end{aligned}$$

Replacing the original terms in (34) by (35), (37) and (38) concludes the proof. □

Each term within the regret decomposition is dealt with differently. We bound the first term via the standard mirror descent analysis, the second term with optimism, the third term with the GOLF regret decomposition and the SEC of Xie et al. (2022), and the fourth term via a modified mirror descent analysis.

The attentive reader will note that the Bellman error is defined as $f - \mathcal{T}f$ in our setup and that of Liu et al. (2023b), and $\mathcal{T}f - f$ in that of Zhong and Zhang (2023).

B.2. Bounding the tracking error

Lemma 9 (Mirror Descent Tracking Error for Algorithm 2)**.** *Let t_k and t_{k+1} be switch times within Algorithm 2, where we use the convention that $\pi^{(t_k)} \propto 1$ is post-policy reset and $\pi^{(t_{k+1})} \not\propto 1$ is pre-policy reset. The tracking error with respect to the optimal policy is then bounded by:*

$$\sum_{t=t_k+1}^{t_{k+1}-1}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle f_h^{(t)}(s_h,\cdot), \pi_h^*(\cdot \mid s_h) - \pi_h^{(t)}(\cdot \mid s_h)\right\rangle\right] \le \eta H^3(t_{k+1} - t_k)/2 + \frac{H\log|\mathcal{A}|}{\eta} + H^2.$$

Proof. Note that for any t such that $t_k + 1 \leq t \leq t_{k+1} - 1$, as we do not reset the policy during these timesteps as the critic does not change, we have

$$\pi_{h+1}^{(t+1)}(\cdot|s') = \frac{\pi_{h+1}^{(t)}(\cdot|s')\exp(\eta f_{h+1}^{(t)}(s',\cdot))}{\sum_{a\in\mathcal{A}}\pi_{h+1}^{(t)}(a|s')\exp(\eta f_{h+1}^{(t)}(s',a))} = Z^{-1}\pi_{h+1}^{(t)}(\cdot|s')\exp(\eta f_{h+1}^{(t)}(s',\cdot)). \tag{39}$$

Rearranging this yields

$$\eta f_{h+1}^{(t)}(s',\cdot) = \log Z + \log \pi_{h+1}^{(t+1)}(\cdot|s') - \log \pi_{h+1}^{(t)}(\cdot|s'),$$

where $\log Z$ is

$$\log Z = \log\left(\sum_{a\in\mathcal{A}}\pi_{h+1}^{(t)}(a|s')\exp(\eta f_{h+1}^{(t)}(s',a))\right).$$

We can now bound, noting that $\sum_{a\in\mathcal{A}}\left(\pi_{h+1}^{*}(\cdot|s') - \pi_{h+1}^{(t+1)}(\cdot|s')\right) = 0$, that

$$\begin{aligned}
&\left\langle \eta f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^{*}(\cdot|s') - \pi_{h+1}^{(t+1)}(\cdot|s')\right\rangle \\
&= \left\langle \log Z + \log \pi_{h+1}^{(t+1)}(\cdot|s') - \log \pi_{h+1}^{(t)}(\cdot|s'), \pi_{h+1}^{*}(\cdot|s') - \pi_{h+1}^{(t+1)}(\cdot|s')\right\rangle \\
&= \left\langle \log \pi_{h+1}^{(t+1)}(\cdot|s') - \log \pi_{h+1}^{(t)}(\cdot|s'), \pi_{h+1}^{*}(\cdot|s') - \pi_{h+1}^{(t+1)}(\cdot|s')\right\rangle \\
&= \mathrm{KL}\left(\pi_{h+1}^{*}(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^{*}(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^{(t+1)}(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right),
\end{aligned} \tag{40}$$

where the last line follows from Lemma 24 with $\pi_1 = \pi_{h+1}^{*}(\cdot|s')$, $\pi_2 = \pi_{h+1}^{(t)}(\cdot|s')$ and $\pi = \pi_{h+1}^{(t+1)}(\cdot|s')$. So it must hold that

$$\begin{aligned}
&\left\langle \eta f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^{*}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')\right\rangle \\
&= \left\langle \eta f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^{*}(\cdot|s') - \pi_{h+1}^{(t+1)}(\cdot|s')\right\rangle - \left\langle \eta f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^{(t)}(\cdot|s') - \pi_{h+1}^{(t+1)}(\cdot|s')\right\rangle \\
&\leq \mathrm{KL}\left(\pi_{h+1}^{*}(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^{*}(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^{(t+1)}(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right) \\
&\quad + \eta H||\pi_{h+1}^{(t+1)}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')||_1.
\end{aligned} \tag{41}$$

Summing up t and h, we can then derive that

$$\begin{aligned}
&\sum_{t=t_k+1}^{t_{k+1}-1}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^{*}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')\right\rangle\right] \\
&= \frac{1}{\eta}\sum_{t=t_k+1}^{t_{k+1}-1}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle \eta f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^{*}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')\right\rangle\right] \\
&\leq \frac{1}{\eta}\sum_{t=t_k+1}^{t_{k+1}-1}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\Big[\mathrm{KL}\left(\pi_{h+1}^{*}(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^{*}(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right) \\
&\qquad\qquad -\mathrm{KL}\left(\pi_{h+1}^{(t+1)}(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right) + \eta H||\pi_{h+1}^{(t+1)}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')||_1\Big].
\end{aligned} \tag{42}$$

Here, we apply Pinsker's inequality on the last line of (42), it follows that

$$\sum_{t=t_k+1}^{t_{k+1}-1}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^{*}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')\right\rangle\right]$$

$$\leq \frac{1}{\eta}\sum_{t=t_k+1}^{t_{k+1}-1}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\Big[\mathrm{KL}\left(\pi^*_{h+1}(\cdot|s')\;||\;\pi^{(t)}_{h+1}(\cdot|s')\right)-\mathrm{KL}\left(\pi^*_{h+1}(\cdot|s')\;||\;\pi^{(t+1)}_{h+1}(\cdot|s')\right)$$
$$-||\pi^{(t+1)}_{h+1}(\cdot|s')-\pi^{(t)}_{h+1}(\cdot|s')||_1^2/2+\eta H||\pi^{(t+1)}_{h+1}(\cdot|s')-\pi^{(t)}_{h+1}(\cdot|s')||_1\Big]$$
$$\leq \frac{1}{\eta}\sum_{t=t_k+1}^{t_{k+1}-1}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\Big[\mathrm{KL}\left(\pi^*_{h+1}(\cdot|s')\;||\;\pi^{(t)}_{h+1}(\cdot|s')\right)-\mathrm{KL}\left(\pi^*_{h+1}(\cdot|s')\;||\;\pi^{(t+1)}_{h+1}(\cdot|s')\right)+\eta^2H^2/2\Big], \tag{43}$$

where we use the fact that $\max_{x\in\mathbb{R}}\left\{-x^2/2+\eta Hx\right\}=\eta^2H^2/2$ in the last line. Continuing to simplify (43) yields

$$\sum_{t=t_k+1}^{t_{k+1}-1}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle f^{(t)}_{h+1}(s',\cdot),\pi^*_{h+1}(\cdot|s')-\pi^{(t)}_{h+1}(\cdot|s')\right\rangle\right]$$
$$\leq \sum_{t=t_k+1}^{t_{k+1}-1}\sum_{h=1}^{H}\left(\eta H^2/2+\mathbb{E}_{\pi^*}\left[\frac{\mathrm{KL}\left(\pi^*_{h+1}(\cdot|s')\;||\;\pi^{(t)}_{h+1}(\cdot|s')\right)-\mathrm{KL}\left(\pi^*_{h+1}(\cdot|s')\;||\;\pi^{(t+1)}_{h+1}(\cdot|s')\right)}{\eta}\right]\right)$$
$$\leq \eta H^3(t_{k+1}-t_k)/2+\sum_{h=1}^{H}\frac{\mathrm{KL}\left(\pi^*_{h+1}(\cdot|s')\;||\;\pi^{(t_k+1)}_{h+1}(\cdot|s')\right)-\mathrm{KL}\left(\pi^*_{h+1}(\cdot|s')\;||\;\pi^{(t_{k+1})}_{h+1}(\cdot|s')\right)}{\eta}$$
$$\leq \eta H^3(t_{k+1}-t_k)/2+\frac{H\log|\mathcal{A}|}{\eta}+H^2, \tag{44}$$

where the last inequality follows from the fact that the KL-divergence is non-negative as well as the policy reset of setting $\pi^{(t_k)}_h$ back to the uniform policy after the critic update. Note that we use the convention that $\pi^{(t_k)}_h$ is after the reset, and $\pi^{(t_{k+1})}_h$ is before the reset. This means that

$$\frac{1}{|\mathcal{A}|\exp(\eta H)}\leq\frac{\exp(0)}{\sum_{a\in\mathcal{A}}\exp(\eta H)}\leq\pi^{(t_k+1)}_{h+1}(\cdot|s')=\frac{\exp(\eta f^{(t)}_{h+1}(s',\cdot))}{\sum_{a\in\mathcal{A}}\exp(\eta f^{(t)}_{h+1}(s',a))}\leq\frac{\exp(\eta H)}{\sum_{a\in\mathcal{A}}\exp(0)}=\frac{\exp(\eta H)}{|\mathcal{A}|},$$

and so we can conclude that

$$\mathrm{KL}\left(\pi^*_{h+1}(\cdot|s')\;||\;\pi^{(t_k+1)}_{h+1}(\cdot|s')\right)\leq\log\left(\frac{1}{\pi^{(t_k+1)}_{h+1}(\cdot|s')}\right)\leq\log|\mathcal{A}|+\eta H. \tag{45}$$

□

B.3. Asserting optimism

Lemma 10 (Negative Bellman Error For Algorithm 2)**.** *Within Algorithm 2, we have that*

$$-\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[f^{(t)}_h-\mathcal{T}^{\pi^*}_h f^{(t)}_{h+1}\right]\leq 0.$$

Proof. Applying Lemma 14, we note that $f^{(t)}_h\geq\mathcal{T}^{\pi^*}_h f^{(t)}_{h+1}$. The result then follows.

□

B.4. Bounding the Bellman error under the learned policies

We now turn our attention to the Bellman error with respect to $\mathcal{T}_h$ under the current policy's occupancy measure.

Lemma 11 (Sum of Bellman Errors Under Algorithm 2)**.** *Within Algorithm 2, the sum of Bellman errors with respect to $\mathcal{T}$ under the occupancy measure of $\pi^{(t)}$ can be bounded by:*

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)}-\mathcal{T}_h f_{h+1}^{(t)}\right)(s_h,a_h)\right]\le\sqrt{\beta H^4 T\mathsf{SEC}(\mathcal{F},\Pi,T)},$$

where $\beta=\Theta\left(\log\left(HT^2\mathcal{N}_{\mathcal{F},\mathcal{TF}}(1/T)/\delta\right)\right)$.

Proof. We can now perform the same Cauchy-Schwarz and change of measure argument as in Xie et al. (2022) to find that

$$\begin{aligned}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)}-\mathcal{T}_h f_{h+1}^{(t)}\right)(s_h,a_h)\right]&=\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{d_h^{(t)}}[\delta_h^{(t)}]\\&=\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{d_h^{(t)}}[\delta_h^{(t)}]\left(\frac{H^2\vee\sum_{i=1}^{t-1}\mathbb{E}_{d_h^{(i)}}[(\delta_h^{(t)})^2]}{H^2\vee\sum_{i=1}^{t-1}\mathbb{E}_{d_h^{(i)}}[(\delta_h^{(t)})^2]}\right)^{1/2}\\&\le\sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H}\frac{\mathbb{E}_{d_h^{(i)}}[\delta_h^{(t)}]^2}{H^2\vee\sum_{i=1}^{t-1}\mathbb{E}_{d_h^{(i)}}[(\delta_h^{(t)})^2]}}\sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H}H^2\vee\sum_{i=1}^{t-1}\mathbb{E}_{d_h^{(i)}}[(\delta_h^{(t)})^2]}.\end{aligned}\tag{46}$$

Within the last inequality, the first term can be bounded by H times the SEC of Xie et al. (2022), by Definition 2. The second term is bounded by Lemma 14. Therefore, (46) can be bounded as

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)}-\mathcal{T}_h f_{h+1}^{(t)}\right)(s_h,a_h)\right]\le\sqrt{H\mathsf{SEC}(\mathcal{F},\Pi,T)}\sqrt{\beta H^3 T}\le\sqrt{\beta H^4 T\mathsf{SEC}(\mathcal{F},\Pi,T)}.\tag{47}$$

□

Lemma 12 (Greedy Policy Tracking Error For Algorithm 2)**.** *Let t_k and t_{k+1} be switch times within Algorithm 2, where we use the convention that $\pi^{(t_k)}\propto 1$ is post-policy reset and $\pi^{(t_{k+1})}\not\propto 1$ is pre-policy reset. The tracking error with respect to the greedy policy corresponding to the current critic is then bounded by:*

$$\sum_{t=t_k+1}^{t_{k+1}-1}\sum_{h=1}^{H}\mathbb{E}_{d_h^{(t)}}\left[\left\langle f_{h+1}^{(t)}(s',\cdot),\pi_{h+1}^{f^{(t)}}(\cdot|s')-\pi_{h+1}^{(t)}(\cdot|s')\right\rangle\right]\le\eta H^3(t_{k+1}-t_k)/2+\frac{H\log|\mathcal{A}|}{\eta}+H^2.$$

Proof. Again note that for any t such that $t_k+1\le t\le t_{k+1}$, we do not reset the policy during these timesteps as the critic does not change. We therefore have

$$\pi_{h+1}^{(t+1)}(\cdot|s')=\frac{\pi_{h+1}^{(t)}(\cdot|s')\exp(\eta f_{h+1}^{(t)}(s',\cdot))}{\sum_{a\in\mathcal{A}}\pi_{h+1}^{(t)}(a|s')\exp(\eta f_{h+1}^{(t)}(s',a))}=Z_t^{-1}\pi_{h+1}^{(t)}(\cdot|s')\exp(\eta f_{h+1}^{(t)}(s',\cdot)),\tag{48}$$

and rearranging this yields

$$\eta f_{h+1}^{(t)}(s',\cdot)=\log Z_t+\log\pi_{h+1}^{(t+1)}(\cdot|s')-\log\pi_{h+1}^{(t)}(\cdot|s'),$$

where $\log Z_t$ is

$$\log Z_t=\log\left(\sum_{a\in\mathcal{A}}\pi_{h+1}^{(t)}(a|s')\exp(\eta f_{h+1}^{(t)}(s',a))\right)=\log\pi_{h+1}^{(t)}(\cdot|s')-\log\pi_{h+1}^{(t+1)}(\cdot|s')+\eta f_{h+1}^{(t)}(s',\cdot).$$

Noting that $\sum_{a\in\mathcal{A}}\left(\pi_{h+1}(\cdot|s')-\pi'_{h+1}(\cdot|s')\right)=0$ for any two policies $\pi,\pi'\in\Pi$, we can now bound

$$\left\langle\eta f_{h+1}^{(t)}(s',\cdot),\pi_{h+1}^{f^{(t)}}(\cdot|s')-\pi_{h+1}^{(t+1)}(\cdot|s')\right\rangle$$

$$
\begin{aligned}
&= \left\langle \log Z + \log \pi_{h+1}^{(t+1)}(\cdot|s') - \log \pi_{h+1}^{(t)}(\cdot|s'), \pi_{h+1}^{f^{(t)}}(\cdot|s') - \pi_{h+1}^{(t+1)}(\cdot|s') \right\rangle \\
&= \left\langle \log \pi_{h+1}^{(t+1)}(\cdot|s') - \log \pi_{h+1}^{(t)}(\cdot|s'), \pi_{h+1}^{f^{(t)}}(\cdot|s') - \pi_{h+1}^{(t+1)}(\cdot|s') \right\rangle \\
&= \mathrm{KL}\left(\pi_{h+1}^{f^{(t)}}(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^{f^{(t)}}(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^{(t+1)}(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right), \qquad (49)
\end{aligned}
$$

where the last line follows from Lemma 24. So it satisfies that

$$
\begin{aligned}
&\left\langle \eta f_{h+1}^{(t)}(s', \cdot), \pi_{h+1}^{f^{(t)}}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s') \right\rangle \\
&= \left\langle \eta f_{h+1}^{(t)}(s', \cdot), \pi_{h+1}^{f^{(t)}}(\cdot|s') - \pi_{h+1}^{(t+1)}(\cdot|s') \right\rangle - \left\langle \eta f_{h+1}^{(t)}(s', \cdot), \pi_{h+1}^{(t)}(\cdot|s') - \pi_{h+1}^{(t+1)}(\cdot|s') \right\rangle \\
&\le \mathrm{KL}\left(\pi_{h+1}^{f^{(t)}}(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^{f^{(t)}}(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^{(t+1)}(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right) \\
&\quad + \eta H ||\pi_{h+1}^{(t+1)}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')||_1. \qquad (50)
\end{aligned}
$$

Sum up with t and h, we can then derive

$$
\begin{aligned}
&\sum_{t=t_k+1}^{t_{k+1}-1} \sum_{h=1}^{H} \mathbb{E}_{d_h^{(t)}} \left[\left\langle f_{h+1}^{(t)}(s', \cdot), \pi_{h+1}^{f^{(t)}}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s') \right\rangle \right] \\
&= \frac{1}{\eta} \sum_{t=t_k+1}^{t_{k+1}-1} \sum_{h=1}^{H} \mathbb{E}_{d_h^{(t)}} \left[\left\langle \eta f_{h+1}^{(t)}(s', \cdot), \pi_{h+1}^{f^{(t)}}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s') \right\rangle \right] \\
&\le \frac{1}{\eta} \sum_{t=t_k+1}^{t_{k+1}-1} \sum_{h=1}^{H} \mathbb{E}_{d_h^{(t)}} \Big[\mathrm{KL}\left(\pi_{h+1}^{f^{(t)}}(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^{f^{(t)}}(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right) \\
&\qquad\qquad - \mathrm{KL}\left(\pi_{h+1}^{(t+1)}(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right) + \eta H ||\pi_{h+1}^{(t+1)}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')||_1 \Big]. \qquad (51)
\end{aligned}
$$

When applying Pinsker's inequality in the last line of (51), we can show that

$$
\begin{aligned}
&\sum_{t=t_k+1}^{t_{k+1}-1} \sum_{h=1}^{H} \mathbb{E}_{d_h^{(t)}} \left[\left\langle f_{h+1}^{(t)}(s', \cdot), \pi_{h+1}^{f^{(t)}}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s') \right\rangle \right] \\
&\le \frac{1}{\eta} \sum_{t=t_k+1}^{t_{k+1}-1} \sum_{h=1}^{H} \mathbb{E}_{d_h^{(t)}} \Big[\mathrm{KL}\left(\pi_{h+1}^{f^{(t)}}(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^{f^{(t)}}(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right) \\
&\qquad\qquad - ||\pi_{h+1}^{(t+1)}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')||_1^2/2 + \eta H ||\pi_{h+1}^{(t+1)}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')||_1 \Big]. \qquad (52)
\end{aligned}
$$

Here, we use the fact that $\max_{x \in \mathbb{R}} \left\{ -x^2/2 + \eta H x \right\} = \eta^2 H^2/2$, and obtain that

$$
-||\pi_{h+1}^{(t+1)}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')||_1^2/2 + \eta H ||\pi_{h+1}^{(t+1)}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')||_1 \le \eta^2 H^2/2.
$$

We now can continue through the following. Note that $f^{(t_k+1)} = ... f^{(t)} = ... = f^{(t_{k+1})}$ due to rare-switching. With a direct calculation, we obtain that

$$
\begin{aligned}
&\sum_{t=t_k+1}^{t_{k+1}-1} \mathbb{E}_{d_h^{(t)}} \left[\frac{\mathrm{KL}\left(\pi_{h+1}^{f^{(t)}}(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^{f^{(t)}}(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right)}{\eta} \right] \\
&= \sum_{t=t_k+1}^{t_{k+1}-1} \mathbb{E}_{d_h^{(t)}} \left[\frac{\mathrm{KL}\left(\pi_{h+1}^{f^{(t_k+1)}}(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^{f^{(t_k+1)}}(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right)}{\eta} \right] \\
&= \mathbb{E}_{d_h^{(t)}} \left[\frac{\mathrm{KL}\left(\pi_{h+1}^{f^{(t_k+1)}}(\cdot|s') \,||\, \pi_{h+1}^{(t_k+1)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^{f^{(t_k+1)}}(\cdot|s') \,||\, \pi_{h+1}^{(t_{k+1})}(\cdot|s')\right)}{\eta} \right]. \qquad (53)
\end{aligned}
$$

Furthermore, there exists some $a^*(s')$ for each s' such that

$$a^*(s) = \arg\max_{a' \in \mathcal{A}} f_{h+1}^{(t)}(s', a') \text{ for all } t_k + 1 \le t \le t_{k+1},$$

hence that $\pi_h^f(a'|s) = \mathbb{1}\left(a' = \arg\max_{a \in \mathcal{A}} f_h(s,a)\right)$. This yields

$$\begin{aligned}
&\mathbb{E}_{d_h^{(t)}}\left[\frac{\mathrm{KL}\left(\pi_{h+1}^{f^{(t_k+1)}}(\cdot|s') \,||\, \pi_{h+1}^{(t_k+1)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^{f^{(t_k+1)}}(\cdot|s') \,||\, \pi_{h+1}^{(t_{k+1})}(\cdot|s')\right)}{\eta}\right] \\
&= \frac{1}{\eta}\mathbb{E}_{d_h^{(t)}}\left[1 \cdot \log\left(\frac{1}{\pi_{h+1}^{(t_k+1)}(a^*(s')|s')}\right) - 1 \cdot \log\left(\frac{1}{\pi_{h+1}^{(t_{k+1})}(a^*(s')|s')}\right)\right] \\
&= \frac{1}{\eta}\mathbb{E}_{d_h^{(t)}}\left[\log\left(\frac{\pi_{h+1}^{(t_{k+1})}(a^*(s')|s')}{\pi_{h+1}^{(t_k+1)}(a^*(s')|s')}\right)\right]. \qquad (54)
\end{aligned}$$

For this term, we note that $\pi_h^{(t_k)}$ is back to the uniform policy after the critic update. Note that we use the convention that $\pi_h^{(t_k)}$ is before the reset, and $\pi_h^{(t_k+1)}$ is after the reset. This means that

$$\frac{1}{|\mathcal{A}|\exp(\eta H)} \le \frac{\exp(0)}{\sum_{a\in\mathcal{A}}\exp(\eta H)} \le \pi_{h+1}^{(t_k+1)}(\cdot|s') = \frac{\exp(\eta f_{h+1}^{(t)}(s',\cdot))}{\sum_{a\in\mathcal{A}}\exp(\eta f_{h+1}^{(t)}(s',a))} \le \frac{\exp(\eta H)}{\sum_{a\in\mathcal{A}}\exp(0)} = \frac{\exp(\eta H)}{|\mathcal{A}|},$$

and so we can conclude that

$$\begin{aligned}
&\sum_{t=t_k+1}^{t_{k+1}-1}\mathbb{E}_{d_h^{(t)}}\left[\frac{\mathrm{KL}\left(\pi_{h+1}^{f^{(t)}}(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right) - \mathrm{KL}\left(\pi_{h+1}^{f^{(t)}}(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right)}{\eta}\right] \\
&= \frac{1}{\eta}\log\left(\frac{\pi_{h+1}^{(t_{k+1})}(a^*(s')|s')}{\pi_{h+1}^{(t_k+1)}(a^*(s')|s')}\right) \le \frac{1}{\eta}\log\left(\frac{1}{\pi_{h+1}^{(t_k+1)}(a^*(s')|s')}\right) \le \frac{1}{\eta}\left(\log|\mathcal{A}| + \eta H\right). \qquad (55)
\end{aligned}$$

Therefore, we obtain that

$$\begin{aligned}
\sum_{t=t_k+1}^{t_{k+1}-1}\sum_{h=1}^{H}\mathbb{E}_{d_h^{(t)}}\left[\left\langle f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^{f^{(t)}}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')\right\rangle\right] &\le \frac{1}{\eta}\sum_{t=t_k+1}^{t_{k+1}-1}\sum_{h=1}^{H}(\eta^2H^2/2) + \frac{1}{\eta}\sum_{h=1}^{H}\left(\log|\mathcal{A}| + \eta H\right) \\
&\le \eta H^3(t_{k+1} - t_k)/2 + \frac{H\log|\mathcal{A}|}{\eta} + H^2. \qquad (56)
\end{aligned}$$

□

B.5. Auxiliary lemmas

B.5.1. Bound on rare-switching update frequency

We bound the rare-switching update frequency via an argument similar to that of Xiong et al. (2023).

Lemma 13 (Switching Costs)**.** *Consider a procedure where the critic is updated only when there exists some h such that*

$$\mathcal{L}_h^{(t)}(f_h^{(t)}, f_{h+1}^{(t)}) - \min_{f_h' \in \mathcal{F}_h}\mathcal{L}_h^{(t)}(f_h', f_{h+1}^{(t)}) \ge 5H^2\beta.$$

This performs no more than $N_{\text{updates},h}(T) \le O(d\log(T))$ Q-function updates for each $h \in [H]$, and no more than $N_{\text{updates}}(T) \le O(dH\log(T))$ Q-function updates in total.

Proof. To show this result, we control the number of switches induced by the Q-function class targeting π^* by upper and lower bounding the cumulative squared Bellman error under the observed states and actions. Fix some $h \in [H]$ for now.

For simplicity, write $K_h = N_{\text{updates},h}(T)$ for the total number of updates, and $t_{1,h}, ..., t_{K_h,h}$ the update times for $f_h^{(t)}$, with $t_{0,h} = 0$. By definition, at every $t_{k,h}$,

$$\mathcal{L}_h^{(t_{k,h})}\left(f_h^{(t_{k,h})}, f_{h+1}^{(t_{k,h})}\right) - \min_{f_h' \in \mathcal{F}_h} \mathcal{L}_h^{(t_{k,h})}\left(f_h', f_{h+1}^{(t_{k,h})}\right) \geq 5H^2\beta. \tag{57}$$

An application of Lemma 17 yields

$$0 \leq \mathcal{L}_h^{(t_{k,h})}\left(\mathcal{T}_h f_{h+1}^{(t_{k,h})}, f_{h+1}^{(t_{k,h})}\right) - \min_{f_h' \in \mathcal{F}_h} \mathcal{L}_h^{(t_{k,h})}\left(f_h', f_{h+1}^{(t_{k,h})}\right) \leq H^2\beta,$$

$$\min_{f_h' \in \mathcal{F}_h} \mathcal{L}_h^{(t_{k,h})}\left(f_h', f_{h+1}^{(t_{k,h})}\right) \geq \mathcal{L}_h^{(t_{k,h})}\left(\mathcal{T}_h f_{h+1}^{(t_{k,h})}, f_{h+1}^{(t_{k,h})}\right) - H^2\beta,$$

$$\mathcal{L}_h^{(t_{k,h})}\left(f_h^{(t_{k,h})}, f_{h+1}^{(t_{k,h})}\right) - \mathcal{L}_h^{(t_{k,h})}\left(\mathcal{T}_h f_{h+1}^{(t_{k,h})}, f_{h+1}^{(t_{k,h})}\right) \geq 4H^2\beta.$$

From the above, one can obtain

$$\begin{aligned}
&\mathcal{L}_h^{(t_{k-1,h}+1:t_{k,h})}\left(f_h^{(t_{k,h})}, f_{h+1}^{(t_{k,h})}\right) - \mathcal{L}_h^{(t_{k-1,h}+1:t_{k,h})}\left(\mathcal{T}_h f_{h+1}^{(t_{k,h})}, f_{h+1}^{(t_{k,h})}\right) \\
&= \mathcal{L}_h^{(t_{k-1,h}+1:t_{k,h})}\left(f_h^{(t_{k-1,h}+1)}, f_{h+1}^{(t_{k-1,h}+1)}\right) - \mathcal{L}_h^{(t_{k-1,h}+1:t_{k,h})}\left(\mathcal{T}_h f_{h+1}^{(t_{k-1,h}+1)}, f_{h+1}^{(t_{k-1,h}+1)}\right) \\
&= \mathcal{L}_h^{(t_{k,h})}\left(f_h^{(t_{k-1,h}+1)}, f_{h+1}^{(t_{k-1,h}+1)}\right) - \mathcal{L}_h^{(t_{k,h})}\left(\mathcal{T}_h f_{h+1}^{(t_{k-1,h}+1)}, f_{h+1}^{(t_{k-1,h}+1)}\right) \\
&\quad - \left(\mathcal{L}_h^{(t_{k-1,h})}\left(f_h^{(t_{k-1,h}+1)}, f_{h+1}^{(t_{k-1,h}+1)}\right) + \mathcal{L}_h^{(t_{k-1,h})}\left(\mathcal{T}_h f_{h+1}^{(t_{k-1,h}+1)}, f_{h+1}^{(t_{k-1,h}+1)}\right)\right) \\
&= \mathcal{L}_h^{(t_{k,h})}\left(f_h^{(t_{k,h})}, f_{h+1}^{(t_{k,h})}\right) - \mathcal{L}_h^{(t_{k,h})}\left(\mathcal{T}_h f_{h+1}^{(t_{k,h})}, f_{h+1}^{(t_{k,h})}\right) \\
&\quad - \left(\mathcal{L}_h^{(t_{k-1,h})}\left(f_h^{(t_{k-1,h}+1)}, f_{h+1}^{(t_{k-1,h}+1)}\right) + \mathcal{L}_h^{(t_{k-1,h})}\left(\mathcal{T}_h f_{h+1}^{(t_{k-1,h}+1)}, f_{h+1}^{(t_{k-1,h}+1)}\right)\right) \\
&\geq 4H^2\beta - H^2\beta = 3H^2\beta.
\end{aligned} \tag{58}$$

Therefore, for any t such that $t_{k-1,h} < t \leq t_{k,h}$, this argument and noting that $f_h^{(t_{k-1,h}+1)} = ... = f_h^{(t)} = ... = f_h^{(t_{k,h})}$ yields

$$\mathcal{L}_h^{(t_{k-1,h}+1:t_{k,h})}\left(f_h^{(t)}, f_{h+1}^{(t)}\right) - \mathcal{L}_h^{(t_{k-1,h}+1:t_{k,h})}\left(\mathcal{T}_h f_{h+1}^{(t)}, f_{h+1}^{(t)}\right) \geq 3H^2\beta. \tag{59}$$

An application of Lemma 19 while noting that $f_h^{(t_{k-1,h}+1)} = ... = f_h^{(t)} = ... = f_h^{(t_{k,h})}$ leads to

$$\sum_{i=t_{k-1,h}+1}^{t_{k,h}} \left(f_h^{(i)} - \mathcal{T}_h f_{h+1}^{(i)}\right)^2 (s_h^{(i)}, a_h^{(i)}) = \sum_{i=t_{k-1,h}+1}^{t_{k,h}} \left(f_h^{(t_{k,h})} - \mathcal{T}_h f_{h+1}^{(t_{k,h})}\right)^2 (s_h^{(i)}, a_h^{(i)}) \geq 2H^2\beta. \tag{60}$$

Now summing over all $t_{1,h}, ..., t_{K,h}$ yields

$$\sum_{t=1}^{T} \left(f_h^{(t)} - \mathcal{T}_h f_{h+1}^{(t)}\right)^2 (s_h^{(t)}, a_h^{(t)}) = \sum_{k=1}^{K_h} \sum_{i=t_{k-1,h}+1}^{t_{k,h}} \left(f_h^{(i)} - \mathcal{T}_h f_{h+1}^{(i)}\right)^2 (s_h^{(i)}, a_h^{(i)}) \geq 2(K_h - 1)H^2\beta. \tag{61}$$

By Lemma 14, we have that

$$\sum_{i=1}^{t-1} \left(f_h^{(t)} - \mathcal{T}_h f_{h+1}^{(t)}\right)^2 (s_h^{(i)}, a_h^{(i)}) \leq O(H^2\beta). \tag{62}$$

Invoking the squared distributional Bellman eluder dimension definition, as in (Xiong et al., 2023), yields

$$\sum_{t=1}^{T} \left(f_h^{(t)} - \mathcal{T}_h f_{h+1}^{(t)}\right)^2 (s_h^{(t)}, a_h^{(t)}) \leq O(dH^2\beta \log T). \tag{63}$$

So we have established that

$$2(K_h - 1)H^2\beta \leq \sum_{t=1}^{T} \left(f_h^{(t)} - \mathcal{T}_h f_{h+1}^{(t)}\right)^2 (s_h^{(t)}, a_h^{(t)}) \leq O(dH^2\beta \log T). \tag{64}$$

The number of updates for each h must therefore be bounded as

$$K_h \leq d \log T, \text{ yielding } N_{\text{switch}}(T) \leq dH \log T.$$

□

B.5.2. Showing optimism for critics targeting Q^*

The following lemma applies to Algorithms 2 and 4. As Algorithm 2 is optimistic, both properties apply to it, while only the second applies to Algorithm 4.

Lemma 14 (Optimism and in-sample error control for critics targeting Q^*)**.** *With probability at least $1-\delta$, for all $t \in [T]$, we have that for all $h = 1, \ldots, H$, an optimistic critic targeting Q^* in the same way as defined in Algorithm 2 achieves*

$$\textit{(i) } \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} \in \mathcal{F}_h^{(t)}, \textit{ and } f_h^{(t)} \geq \mathcal{T}_h^{\pi} f_{h+1}^{(t)} \textit{ for all } \pi,$$

Similarly, a critic targeting Q^* as in Algorithms 2 and 4 achieves

$$\textit{(ii) } \sum_{i=1}^{t-1} \mathbb{E}_{d^{\pi^{(i)}}} \left[\left(f_h^{(t)} - \mathcal{T}_h f_{h+1}^{(t)}\right)^2\right] \leq O\left(H^2\beta\right),$$

by choosing $\beta = c_1 \left(\log\left[HT\mathcal{N}_{\mathcal{F},\mathcal{TF}}(\rho)/\delta\right]\right)$ for some constant c_1.

Proof. By Lemma 17 applied to the greedy policy $\pi^{f^{(t)}}$, for any $h \in [H]$ and $t \in [T]$, we have with probability $1-\delta$ that

$$0 \leq \mathcal{L}_h^{(t)}(\mathcal{T}_h f_{h+1}^{(t)}, f_{h+1}^{(t)}) - \min_{f_h' \in \mathcal{F}_h} \mathcal{L}_h^{(t)}(f_h', f_{h+1}^{(t)}) \leq H^2\beta. \tag{65}$$

We construct the confidence sets only at timesteps t_k where switches occur:

$$\mathcal{F}_h^{(t_{\text{last}})} := \left\{ f \in \mathcal{F} : \mathcal{L}_h^{(t_{\text{last}})}(f_h, f_{h+1}) - \min_{f_h' \in \mathcal{F}_h} \mathcal{L}_h^{(t_{\text{last}})}(f_h', f_{h+1}) \leq H^2\beta \ \ \forall h \in [H] \right\}.$$

We can now show the first result. As we defined $f_h^{(t)}(s,a) := \operatorname{argmax}_{f_h \in \mathcal{F}_h^{(t_{\text{last}})}} f_h(s,a)$, for all π:

$$\mathcal{T}_h^{\pi} f_{h+1}^{(t)}(s,a) = r_h(s,a) + \mathbb{E}_{s'}\left[f_{h+1}^{(t)}(s', \pi_{h+1}(s'))\right] \leq r_h(s,a) + \mathbb{E}_{s'}\left[\max_{a' \in \mathcal{A}} f_{h+1}^{(t)}(s',a')\right] = \mathcal{T}_h f_{h+1}^{(t)}(s,a). \tag{66}$$

We further note that $\mathcal{T}_h f_{h+1}^{(t)}(s,a) \leq f_h^{(t)}(s,a)$, as $\mathcal{T}_h f_{h+1}^{(t)} \in \mathcal{F}_h^{(t_{\text{last}})}$ and by the definition of $f_h^{(t)}(s,a)$.

The second result can now be shown using by Lemma 18 using a similar argument to the proof of Theorem 1 in Xiong et al. (2023), which in turn takes inspiration from the proofs of Lemmas 39 and 40 in Jin et al. (2021). We elaborate accordingly.

Consider two cases, one where we perform an update at episode $t-1$ and one where we do not. If we perform an update at episode $t-1$, then by the choice of $f^{(t)}$ to be near-optimal (in fact, with Algorithm 4, it is optimal and this is zero) it must hold that

$$\mathcal{L}_h^{(t-1)}\left(f_h^{(t)}, f_{h+1}^{(t)}\right) - \min_{f_h' \in \mathcal{F}_h} \mathcal{L}_h^{(t-1)}\left(f_h', f_{h+1}^{(t)}\right) \leq 5H^2\beta,$$

and by Lemma 17 it must also hold that

$$0 \leq \mathcal{L}_h^{(t-1)}\left(\mathcal{T}_h f_h^{(t)}, f_{h+1}^{(t)}\right) - \min_{f_h' \in \mathcal{F}_h} \mathcal{L}_h^{(t-1)}\left(f_h', f_{h+1}^{(t)}\right) \leq H^2\beta.$$

One can then see that

$$\mathcal{L}_h^{(t-1)}\left(f_h^{(t)}, f_{h+1}^{(t)}\right) - \mathcal{L}_h^{(t-1)}\left(\mathcal{T}_h f_h^{(t)}, f_{h+1}^{(t)}\right) \leq 6H^2\beta. \tag{67}$$

The same holds for the other case where we do not perform an update at episode $t-1$. Observe that because we did not perform an update,

$$\mathcal{L}_h^{(t-1)}\left(f_h^{(t-1)}, f_{h+1}^{(t-1)}\right) - \min_{f_h' \in \mathcal{F}_h} \mathcal{L}_h^{(t-1)}\left(f_h', f_{h+1}^{(t-1)}\right) \leq 5H^2\beta.$$

From Lemma 17, it also holds that

$$\mathcal{L}_h^{(t-1)}\left(\mathcal{T}_h f_h^{(t-1)}, f_{h+1}^{(t-1)}\right) - \min_{f_h' \in \mathcal{F}_h} \mathcal{L}_h^{(t-1)}\left(f_h', f_{h+1}^{(t-1)}\right) \leq H^2\beta.$$

Putting the above two statements together and using that $f^{(t)} = f^{(t-1)}$ again yields

$$\mathcal{L}_h^{(t-1)}\left(f_h^{(t)}, f_{h+1}^{(t)}\right) - \mathcal{L}_h^{(t-1)}\left(\mathcal{T}_h f_h^{(t)}, f_{h+1}^{(t)}\right) = \mathcal{L}_h^{(t-1)}\left(f_h^{(t-1)}, f_{h+1}^{(t-1)}\right) - \mathcal{L}_h^{(t-1)}\left(\mathcal{T}_h f_h^{(t-1)}, f_{h+1}^{(t-1)}\right) \leq 6H^2\beta. \tag{68}$$

An application of Lemma 18 to both cases then yields in either case that

$$\sum_{i=1}^{t-1}\left(f_h^{(t)} - \mathcal{T}_h f_{h+1}^{(t)}\right)^2 (s_h^{(i)}, a_h^{(i)}) \leq 7H^2\beta, \tag{69}$$

and also that

$$\sum_{i=1}^{t-1}\mathbb{E}_{d_h^{(i)}}\left[\left(f_h^{(t)} - \mathcal{T}_h f_{h+1}^{(t)}\right)^2\right] \leq 7H^2\beta. \tag{70}$$

□

C. Extension of Algorithm 2 to hybrid RL

Before we proceed, it is useful to introduce the single-policy concentrability coefficient tweaked from that of Zhan et al. (2022) and partial all-policy concentrability coefficient from Tan and Xu (2024):

Definition 4 (Single-Policy Concentrability Coefficient)**.**

$$c_{\text{off}}^*(\mathcal{F}, \Pi) = \max_{h\in[H]} \sup_{f\in\mathcal{F}} \sup_{\pi\in\Pi} \frac{\mathbb{E}_{\pi^*}\left[f_h - \mathcal{T}_h^\pi f_{h+1}\right]^2}{\mathbb{E}_\mu[(f_h - \mathcal{T}_h^\pi f_{h+1})^2]}, \quad c_{\text{off}}^*(\mathcal{F}) = \max_{h\in[H]} \sup_{f\in\mathcal{F}} \frac{\mathbb{E}_{\pi^*}\left[f_h - \mathcal{T}_h f_{h+1}\right]^2}{\mathbb{E}_\mu[(f_h - \mathcal{T}_h f_{h+1})^2]}.$$

Definition 5 (Partial All-Policy Concentrability Coefficient)**.** *For a function class $\mathcal{F}$ and a partition on the state-action space $\mathcal{X} = \mathcal{S}\times\mathcal{A}\times[H]$, where we denote the offline and online partitions by $\mathcal{X}_{\text{off}}$ and $\mathcal{X}_{\text{on}}$, respectively, the partial all-policy concentrability coefficient is given by:*

$$c_{\text{off}}(\mathcal{F}\mathbb{1}_{\mathcal{X}_{\text{off}}}) \coloneqq \max_h \sup_{f\in\mathcal{F}} \frac{\|(f_h - \mathcal{T}_h f_{h+1})\mathbb{1}_{(\cdot,h)\in\mathcal{X}_{\text{off}}}\|_{2,d_h^\pi}^2}{\|(f_h - \mathcal{T}_h f_{h+1})\mathbb{1}_{(\cdot,h)\in\mathcal{X}_{\text{off}}}\|_{2,\mu_h}^2},$$

where $\mathbb{1}_{\mathcal{X}_{\text{off}}}$ denotes the indicator variable for whether $s,a \in \mathcal{X}_{\text{on}}$

We now present Algorithm 3, Algorithm 4 and Algorithm 5, which are the modified versions of Algorithm 2 for hybrid RL.

Algorithm 3 (NOAH-π) targets $\pi^{(t)}$, and follows the very natural procedure of performing a critic update via Fitted Q-Evaluation (FQE) (Munos and Szepesvári, 2008) and an actor update in every episode. It therefore requires closure of the critic function class under truncated sums as in Definition 3 to control the growth of the policy class as in Lemma 2. Algorithm 4 (NOAH-$*$), like NORA in Algorithm 2, circumvents this by targeting π^* and performing a rare-switching critic update. Both algorithms are fully off-policy, utilizing offline data and all collected online data without throwing any away.

Algorithm 3 Non-Optimistic Actor-critic with Hybrid RL targeting $\pi^{(t)}$ (NOAH-π)

1: **Input:** Function class $\mathcal{F}$.
2: **Initialize:** $\mathcal{F}^{(0)} \leftarrow \mathcal{F}$, $\mathcal{D}_h^{(0)} \leftarrow \emptyset, \forall h \in [H]$, $\eta = \Theta(\sqrt{\log|\mathcal{A}|H^{-2}T^{-1}})$, $\pi^{(1)} \propto 1$, confidence width $\beta = \Theta(\log(HT^2\mathcal{N}_{\mathcal{F},\mathcal{TF}}(1/T)/\delta))$.
3: **for** episode $t = 1, 2, \ldots, T$ **do**
4: Play policy $\pi^{(t)}$ for one episode, obtain trajectory, update dataset $\mathcal{D}_h^{(t)}$.
5: Compute critic $f^{(t+1)}$ targeting $\pi^{(t)}$ via FQE:

$$f^{(t)} \leftarrow \arg\min_{f\in\mathcal{F}} \mathcal{L}_h^{(t,\pi^{(t)})}(f_h, f_{h+1}) \text{ for } h = H-1, ..., 1,$$
$$\mathcal{L}_h^{(t,\pi^{(t)})}(f, f') \leftarrow \sum_{(s,a,r,s')\in\mathcal{D}_h^{(t)}\cup\mathcal{D}_{\text{off}}} \Big(f(s,a) - r - f'(s', \pi_{h+1}^{(t)}(s'))\Big)^2.$$

6: Update $\pi_h^{(t+1)}(a|s) \propto \pi_h^{(t)}(a|s)\exp(\eta f_h^{(t)}(s,a))$.
7: **end for**

D. Proofs for Regret Guarantees of Hybrid RL

D.1. Proofs for Theorem 3, Algorithm 3

We start with the same regret decomposition as in Lemma 3:

$$\begin{aligned}
\mathrm{Reg}(T) =& \sum_{t=1}^{T}\left(V_1^*(s_1^{(t)}) - V_1^{\pi^{(t)}}(s_1^{(t)})\right) \\
=& \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle f_h^{(t)}(s_h,\cdot), \pi_h^*(\cdot\mid s_h) - \pi_h^{(t)}(\cdot\mid s_h)\right\rangle\right] - \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^*} f_{h+1}^{(t)}\right)(s_h,a_h)\right] \\
&+ \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}\right)(s_h,a_h)\right]. \qquad (71)
\end{aligned}$$

We control the first term with the same argument as Theorem 1, by using Lemma 4. Controlling the third term follows by the same argument as in Lemma 6. It remains to tackle the second term, which we bound with the offline data.

First, we decompose the last term of (71) as

$$-\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[f_h^{(t)} - \mathcal{T}_h^{\pi^*} f_{h+1}^{(t)}\right] = \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} - f_h^{(t)}\right] + \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\mathcal{T}_h^{\pi^*} f_{h+1}^{(t)} - \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}\right]. \qquad (72)$$

The latter term of (72) is bounded as:

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\mathcal{T}_h^{\pi^*} f_{h+1}^{(t)} - \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}\right] \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[r_h(s_h,a_h) + f_{h+1}^{(t)}\left(s_{h+1}, \pi_{h+1}^*(s_{h+1})\right) - r_h(s_h,a_h) - f_{h+1}^{(t)}\left(s_{h+1}, \pi_{h+1}^{(t)}(s_{h+1})\right)\right] \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[f_{h+1}^{(t)}\left(s_{h+1}, \pi_{h+1}^*(s_{h+1})\right) - f_{h+1}^{(t)}\left(s_{h+1}, \pi_{h+1}^{(t)}(s_{h+1})\right)\right] \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle f_{h+1}^{(t)}(s_{h+1},\cdot), \pi_{h+1}^*(\cdot|s_{h+1}) - \pi_{h+1}^{(t)}(\cdot|s_{h+1})\right\rangle\right]. \qquad (73)
\end{aligned}$$

Algorithm 4 Non-Optimistic Actor-critic with Hybrid RL targeting π^* (NOAH-$*$)

Input: Function class $\mathcal{F}$.
Initialize: $\mathcal{F}^{(0)} \leftarrow \mathcal{F}$, $\mathcal{D}_h^{(0)} \leftarrow \emptyset, \forall h \in [H]$, $\eta = \Theta(\sqrt{d \log T \log |\mathcal{A}| H^{-1} T^{-1}})$, $\pi^{(1)} \propto 1$, confidence width $\beta = \Theta(\log(HT^2 \mathcal{N}_{\mathcal{F},\mathcal{T}\mathcal{F}}(1/T)/\delta))$.
for episode $t = 1, 2, \ldots, T$ **do**
  Play policy $\pi^{(t)}$ for one episode, obtain trajectory, update dataset $\mathcal{D}_h^{(t)}$.
  if $\mathcal{L}_h^{(t)}(f_h^{(t)}, f_{h+1}^{(t)}) \geq \min_{f'_h \in \mathcal{F}_h} \mathcal{L}_h^{(t)}(f'_h, f_{h+1}^{(t)}) + 5H^2\beta$ for some h **then**
    Compute critic $f^{(t+1)}$ via FQE:

$$f^{(t)} \leftarrow \arg\min_{f \in \mathcal{F}} \mathcal{L}_h^{(t)}(f_h, f_{h+1}) \text{ for } h = H-1, ..., 1,$$
$$\mathcal{L}_h^{(t)}(f, f') \leftarrow \sum_{(s,a,r,s') \in \mathcal{D}_h^{(t)} \cup \mathcal{D}_{\text{off}}} \left(f(s,a) - r - \max_{a' \in \mathcal{A}} f'(s', a') \right)^2.$$

    Reset policy $\pi^{(t)} \propto 1$.
    Set $t_{\text{last}} \leftarrow t$, $N_{\text{updates}}^{(t)} \leftarrow N_{\text{updates}}^{(t-1)} + 1$.
  else
    Set $N_{\text{updates}}^{(t)} \leftarrow N_{\text{updates}}^{(t-1)}$, $f^{(t+1)} \leftarrow f^{(t)}$.
  end if
  Update $\pi_h^{(t+1)}(a|s) \propto \pi_h^{(t)}(a|s) \exp(\eta f_h^{(t)}(s,a))$.
end for

This yields what is essentially a copy of the first term of (71). For the former term, we use a similar argument to that of Theorem 4. Concretely, as long as $-f \in \mathcal{F}$ for all $f \in \mathcal{F}$,

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\left(\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} - f_h^{(t)}\right)(s_h, a_h)\right] \\
&\leq \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\left(\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} - f_h^{(t)}\right)\right] \left(\frac{N_{\text{off}} \mathbb{E}_\mu\left[\left(\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} - f_h^{(t)}\right)^2\right] + \sum_{i=1}^{t-1} \mathbb{E}_{d_h^{(i)}}\left[\left(\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} - f_h^{(t)}\right)^2\right]}{N_{\text{off}} \mathbb{E}_\mu\left[\left(\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} - f_h^{(t)}\right)^2\right] + \sum_{i=1}^{t-1} \mathbb{E}_{d_h^{(i)}}\left[\left(\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} - f_h^{(t)}\right)^2\right]} \right)^{1/2} \\
&\leq \sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H} \frac{\mathbb{E}_{\pi^*}\left[\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} - f_h^{(t)}\right]^2}{N_{\text{off}} \mathbb{E}_\mu\left[\left(\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} - f_h^{(t)}\right)^2\right] + \sum_{i=1}^{t-1} \mathbb{E}_{d_h^{(i)}}\left[\left(\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} - f_h^{(t)}\right)^2\right]}} \\
&\quad \cdot \sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H} \left(N_{\text{off}} \mathbb{E}_\mu\left[\left(\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} - f_h^{(t)}\right)^2\right] + \sum_{i=1}^{t-1} \mathbb{E}_{d_h^{(i)}}\left[\left(\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} - f_h^{(t)}\right)^2\right] \right)} \\
&\leq \sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H} \frac{\mathbb{E}_{\pi^*}\left[\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} - f_h^{(t)}\right]^2}{N_{\text{off}} \mathbb{E}_\mu\left[\left(\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} - f_h^{(t)}\right)^2\right]}} \\
&\quad \cdot \sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H} \left(N_{\text{off}} \mathbb{E}_\mu\left[\left(\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} - f_h^{(t)}\right)^2\right] + \sum_{i=1}^{t-1} \mathbb{E}_{d_h^{(i)}}\left[\left(\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} - f_h^{(t)}\right)^2\right] \right)} \\
&\leq \sqrt{c^*_{\text{off}}(\mathcal{F}, \Pi) HT / N_{\text{off}}} \sqrt{\beta H^3 T} \\
&\leq \sqrt{H^4 \beta c^*_{\text{off}}(\mathcal{F}, \Pi) T^2 / N_{\text{off}}},
\end{aligned} \tag{74}$$

Algorithm 5 No-regret Optimistic Rare-switching Actor-critic (NORA) for Hybrid RL

Input: Offline dataset $\mathcal{D}_{\text{off}}$ which can be the empty set, samples sizes T, N_{off}, function class $\mathcal{F}$ and confidence width $\beta > 0$
Initialize: $\mathcal{F}^{(0)} \leftarrow \mathcal{F}$, $\mathcal{D}_h^{(0)} \leftarrow \emptyset, \forall h \in [H]$, $\eta = \Theta(\sqrt{d \log T \log |\mathcal{A}| H^{-1} T^{-1}})$, $\pi^{(1)} \propto 1$.
for episode $t = 1, 2, \ldots, T$ **do**
  Select critic $f_h^{(t)}(s,a) := \operatorname{argmax}_{f_h \in \mathcal{F}_h^{(t_{\text{last}})}} f_h(s,a)$ for all s, a.
  Play policy $\pi^{(t)}$ for one episode and obtain trajectory $(s_1^{(t)}, a_1^{(t)}, r_1^{(t)}), \ldots, (s_H^{(t)}, a_H^{(t)}, r_H^{(t)})$.
  Update dataset $\mathcal{D}_h^{(t)} \leftarrow \mathcal{D}_h^{(t-1)} \cup \{(s_h^{(t)}, a_h^{(t)}, r_h^{(t)}, s_{h+1}^{(t)})\}, \forall h \in [H]$.
  if there exists some h such that $\mathcal{L}_h^{(t)}(f_h^{(t)}, f_{h+1}^{(t)}) - \min_{f_h' \in \mathcal{F}_h} \mathcal{L}_h^{(t)}(f_h', f_{h+1}^{(t)}) \geq 5H^2\beta$ **then**
    Compute confidence set $\mathcal{F}^{(t)}$:

$$\mathcal{F}^{(t)} \leftarrow \left\{ f \in \mathcal{F} : \mathcal{L}_h^{(t)}(f_h, f_{h+1}) - \min_{f_h' \in \mathcal{F}_h} \mathcal{L}_h^{(t)}(f_h', f_{h+1}) \leq H^2\beta \quad \forall h \in [H] \right\},$$

$$\text{where } \mathcal{L}_h^{(t)}(f, f') := \sum_{(s,a,r,s') \in \mathcal{D}_h^{(t)} \cup \mathcal{D}_{\text{off},h}} \left(f(s,a) - r - \max_{a' \in \mathcal{A}} f'(s', a') \right)^2, \forall f \in \mathcal{F}_h, f' \in \mathcal{F}_{h+1}.$$

    Reset policy $\pi^{(t)} \propto 1$.
    Set $t_{\text{last}} := t$, increment number of updates $N_{\text{updates}}^{(t)} := N_{\text{updates}}^{(t-1)} + 1$.
  else
    Set $N_{\text{updates}}^{(t)} := N_{\text{updates}}^{(t-1)}$, $\mathcal{F}^{(t)} := \mathcal{F}^{(t-1)}$.
  end if
  Select policy $\pi_h^{(t+1)} \propto \pi_h^{(t)} \exp(\eta f_h^{(t)})$.
end for

where for the penultimate line, the bound for the first term follows directly from the Definition 4 on single-policy conentrability coefficient, and the second bound follows directly from Lemma 15. Note that the argument is similar to that of Theorem 4, with the exception that we can use the single-policy concentrability coefficient as the density ratio we need to bound is

$$\frac{\mathbb{E}_{\pi^*}\left[\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} - f_h^{(t)}\right]^2}{\mathbb{E}_\mu\left[\left(\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} - f_h^{(t)}\right)^2\right]} \leq \max_{h \in [H]} \sup_{f \in \mathcal{F}} \sup_{\pi \in \Pi} \frac{\mathbb{E}_{\pi^*}\left[f_h - \mathcal{T}_h^\pi f_{h+1}\right]^2}{\mathbb{E}_\mu[(f_h - \mathcal{T}_h f_{h+1})^2]} = c_{\text{off}}^*(\mathcal{F}, \Pi), \tag{75}$$

where again the first inequality holds as long as $-f \in \mathcal{F}$ for all $f \in \mathcal{F}$.

D.2. Proofs for Theorem 3, Algorithm 4

We start with the same regret decomposition in Lemma 8:

$$\begin{aligned}
\operatorname{Reg}(T) &= \sum_{t=1}^T \left(V_1^*(s_1^{(t)}) - V_1^{\pi^{(t)}}(s_1^{(t)}) \right) \\
&= \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\left\langle f_h^{(t)}(s_h, \cdot), \pi_h^*(\cdot \mid s_h) - \pi_h^{(t)}(\cdot \mid s_h) \right\rangle \right] - \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\left(f_h^{(t)} - \mathcal{T}_h^{\pi^*} f_{h+1}^{(t)} \right)(s_h, a_h) \right] \\
&\quad + \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)} - \mathcal{T}_h f_{h+1}^{(t)} \right)(s_h, a_h) \right] + \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_{\pi^{(t)}}\left[\left\langle f_{h+1}^{(t)}(s_{h+1}, \cdot), \pi_{h+1}^{f^{(t)}}(\cdot | s_{h+1}) - \pi_{h+1}^{(t)}(\cdot | s_{h+1}) \right\rangle \right] \\
&= \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\left\langle f_h^{(t)}(s_h, \cdot), \pi_h^*(\cdot \mid s_h) - \pi_h^{(t)}(\cdot \mid s_h) \right\rangle \right] \\
&\quad - \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\left(f_h^{(t)} - \mathcal{T}_h f_{h+1}^{(t)} \right)(s_h, a_h) \right] + \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\left\langle f_{h+1}^{(t)}(s_{h+1}, \cdot), \pi_{h+1}^*(\cdot | s_{h+1}) - \pi_{h+1}^{f^{(t)}}(\cdot | s_{h+1}) \right\rangle \right]
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)}-\mathcal{T}_h f_{h+1}^{(t)}\right)(s_h,a_h)\right]+\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{(t)}}\left[\left\langle f_{h+1}^{(t)}(s_{h+1},\cdot),\pi_{h+1}^{f^{(t)}}(\cdot|s_{h+1})-\pi_{h+1}^{(t)}(\cdot|s_{h+1})\right\rangle\right]\\
=&\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle f_h^{(t)}(s_h,\cdot),\pi_h^*(\cdot\mid s_h)-\pi_h^{(t)}(\cdot\mid s_h)\right\rangle\right]\\
&-\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left(f_h^{(t)}-\mathcal{T}_h f_{h+1}^{(t)}\right)(s_h,a_h)\right]+\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle f_{h+1}^{(t)}(s_{h+1},\cdot),\pi_{h+1}^{*}(\cdot|s_{h+1})-\pi_{h+1}^{(t)}(\cdot|s_{h+1})\right\rangle\right]\\
&+\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle f_{h+1}^{(t)}(s_{h+1},\cdot),\pi_{h+1}^{(t)}(\cdot|s_{h+1})-\pi_{h+1}^{f^{(t)}}(\cdot|s_{h+1})\right\rangle\right]\\
&+\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)}-\mathcal{T}_h f_{h+1}^{(t)}\right)(s_h,a_h)\right]+\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{(t)}}\left[\left\langle f_{h+1}^{(t)}(s_{h+1},\cdot),\pi_{h+1}^{f^{(t)}}(\cdot|s_{h+1})-\pi_{h+1}^{(t)}(\cdot|s_{h+1})\right\rangle\right],
\end{aligned}
\tag{76}
$$

where we break up the original second term corresponding to the negative Bellman error under π^* with a similar decomposition as in the proof of Lemma 8. Now, observe that

$$
\sum_{t=1}^{T}\sum_{h=2}^{H}\mathbb{E}_{\pi^*}\left[\left\langle f_h^{(t)}(s_h,\cdot),\pi_h^{(t)}(\cdot|s_h)-\pi_h^{f^{(t)}}(\cdot|s_h)\right\rangle\right]=\sum_{t=1}^{T}\sum_{h=2}^{H}\mathbb{E}_{\pi^*}\left[f_h^{(t)}(s_h,\pi_h^{(t)}(s_h))-\max_{a\in\mathcal{A}}f_h^{(t)}(s_h,a)\right]\le 0. \tag{77}
$$

So the remaining regret decomposition is:

$$
\begin{aligned}
\mathrm{Reg}(T)=&\sum_{t=1}^{T}\left(V_1^*(s_1^{(t)})-V_1^{\pi^{(t)}}(s_1^{(t)})\right)\\
=&\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle f_h^{(t)}(s_h,\cdot),\pi_h^*(\cdot\mid s_h)-\pi_h^{(t)}(\cdot\mid s_h)\right\rangle\right]-\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left(f_h^{(t)}-\mathcal{T}_h f_{h+1}^{(t)}\right)(s_h,a_h)\right]\\
&+\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle f_{h+1}^{(t)}(s_{h+1},\cdot),\pi_{h+1}^{*}(\cdot|s_{h+1})-\pi_{h+1}^{(t)}(\cdot|s_{h+1})\right\rangle\right]\\
&+\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{(t)}}\left[\left(f_h^{(t)}-\mathcal{T}_h f_{h+1}^{(t)}\right)(s_h,a_h)\right]+\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{(t)}}\left[\left\langle f_{h+1}^{(t)}(s_{h+1},\cdot),\pi_{h+1}^{f^{(t)}}(\cdot|s_{h+1})-\pi_{h+1}^{(t)}(\cdot|s_{h+1})\right\rangle\right]
\end{aligned}
\tag{78}
$$

The proof then follows analogously to Theorem 2 for the first, third, fourth and fifth terms. Note that the proofs for Lemmas 17, 18, 19, and Lemma 13 still hold. Intuitively, this is because the first three lemmas deal with the generalization error of the empirical TD loss under the occupancy measure of the current policy, and the switch cost proof depends only on these lemmas and choosing some $f^{(t)}$ with low enough training error. Choosing the minimizer as in Algorithm 4 fulfills this condition.

Unlike the analysis for Algorithm 2, we bound the second term with the offline data here. We use a similar argument to that of Theorem 4. Concretely, as long as $-f\in\mathcal{F}$ for all $f\in\mathcal{F}$,

$$
\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left(\mathcal{T}_h f_{h+1}^{(t)}-f_h^{(t)}\right)(s_h,a_h)\right]\\
&\le\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left(\mathcal{T}_h f_{h+1}^{(t)}-f_h^{(t)}\right)\right]\left(\frac{N_{\mathrm{off}}\mathbb{E}_\mu\left[\left(\mathcal{T}_h f_{h+1}^{(t)}-f_h^{(t)}\right)^2\right]+\sum_{i=1}^{t-1}\mathbb{E}_{d_h^{(i)}}\left[\left(\mathcal{T}_h f_{h+1}^{(t)}-f_h^{(t)}\right)^2\right]}{N_{\mathrm{off}}\mathbb{E}_\mu\left[\left(\mathcal{T}_h f_{h+1}^{(t)}-f_h^{(t)}\right)^2\right]+\sum_{i=1}^{t-1}\mathbb{E}_{d_h^{(i)}}\left[\left(\mathcal{T}_h f_{h+1}^{(t)}-f_h^{(t)}\right)^2\right]}\right)^{1/2}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H}\frac{\mathbb{E}_{\pi^*}\left[\mathcal{T}_h f_{h+1}^{(t)} - f_h^{(t)}\right]^2}{N_{\text{off}}\mathbb{E}_{\mu}\left[\left(\mathcal{T}_h f_{h+1}^{(t)} - f_h^{(t)}\right)^2\right] + \sum_{i=1}^{t-1}\mathbb{E}_{d_h^{(i)}}\left[\left(\mathcal{T}_h f_{h+1}^{(t)} - f_h^{(t)}\right)^2\right]}} \\
&\quad \cdot \sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H}\left(N_{\text{off}}\mathbb{E}_{\mu}\left[\left(\mathcal{T}_h f_{h+1}^{(t)} - f_h^{(t)}\right)^2\right] + \sum_{i=1}^{t-1}\mathbb{E}_{d_h^{(i)}}\left[\left(\mathcal{T}_h f_{h+1}^{(t)} - f_h^{(t)}\right)^2\right]\right)} \\
&\leq \sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H}\frac{\mathbb{E}_{\pi^*}\left[\mathcal{T}_h f_{h+1}^{(t)} - f_h^{(t)}\right]^2}{N_{\text{off}}\mathbb{E}_{\mu}\left[\left(\mathcal{T}_h f_{h+1}^{(t)} - f_h^{(t)}\right)^2\right]}} \\
&\quad \cdot \sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H}\left(N_{\text{off}}\mathbb{E}_{\mu}\left[\left(\mathcal{T}_h f_{h+1}^{(t)} - f_h^{(t)}\right)^2\right] + \sum_{i=1}^{t-1}\mathbb{E}_{d_h^{(i)}}\left[\left(\mathcal{T}_h f_{h+1}^{(t)} - f_h^{(t)}\right)^2\right]\right)} \\
&\leq \sqrt{c^*_{\text{off}}(\mathcal{F})HT/N_{\text{off}}}\sqrt{\beta H^3 T} \\
&\leq \sqrt{H^4\beta c^*_{\text{off}}(\mathcal{F})T^2/N_{\text{off}}},
\end{aligned} \tag{79}
$$

where for the penultimate line, the bound for the first term follows directly from the Definition 4 on single-policy conentrability coefficient, and the second bound follows directly from Lemma 16. Note that the argument is similar to that of Theorem 4, with the exception that we can use the single-policy concentrability coefficient as the density ratio we need to bound is

$$
\frac{\mathbb{E}_{\pi^*}\left[\mathcal{T}_h f_{h+1}^{(t)} - f_h^{(t)}\right]^2}{\mathbb{E}_{\mu}\left[\left(\mathcal{T}_h f_{h+1}^{(t)} - f_h^{(t)}\right)^2\right]} \leq \max_{h\in[H]}\sup_{f\in\mathcal{F}}\frac{\mathbb{E}_{\pi^*}\left[f_h - \mathcal{T}_h f_{h+1}\right]^2}{\mathbb{E}_{\mu}[\left(f_h - \mathcal{T}_h f_{h+1}\right)^2]} = c^*_{\text{off}}(\mathcal{F}), \tag{80}
$$

where again the first inequality holds as long as $-f \in \mathcal{F}$ for all $f \in \mathcal{F}$.

We further note that it is possible to have the critic target $Q^{\pi^{(t)}}$ with this framework, if the sum of truncated critics is a critic. Fortunately, there is no need to deal with a rarely-updating bonus class like in (Sherman et al., 2024; Cassel and Rosenberg, 2024), as we do not need optimism.

D.3. Proofs for Theorem 4

The proof follows analogously to that of the original online case, in line with the analysis and observations of Tan and Xu (2024). The only difference is that we will now bound the Bellman error under the current policy's occupancy measure as

$$
\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{d_h^{(t)}}[\delta_h^{(t)}] = \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{d_h^{(t)}}[\delta_h^{(t)}\mathbb{1}_{\mathcal{X}_{\text{off}}}] + \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{d_h^{(t)}}[\delta_h^{(t)}\mathbb{1}_{\mathcal{X}_{\text{on}}}]. \tag{81}
$$

For the online term, this is bounded with the same Cauchy-Schwarz and change of measure argument as in Xie et al. (2022):

$$
\begin{aligned}
\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{d_h^{(t)}}[\delta_h^{(t)}\mathbb{1}_{\mathcal{X}_{\text{on}}}] &= \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{d_h^{(t)}}[\delta_h^{(t)}\mathbb{1}_{\mathcal{X}_{\text{on}}}]\left(\frac{H^2 \vee \sum_{i=1}^{t-1}\mathbb{E}_{d_h^{(i)}}[(\delta_h^{(t)})^2\mathbb{1}_{\mathcal{X}_{\text{on}}}]}{H^2 \vee \sum_{i=1}^{t-1}\mathbb{E}_{d_h^{(i)}}[(\delta_h^{(t)})^2\mathbb{1}_{\mathcal{X}_{\text{on}}}]}\right)^{1/2} \\
&\leq \sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H}\frac{\mathbb{E}_{d_h^{(i)}}[\delta_h^{(t)}\mathbb{1}_{\mathcal{X}_{\text{on}}}]^2}{H^2 \vee \sum_{i=1}^{t-1}\mathbb{E}_{d_h^{(i)}}[(\delta_h^{(t)})^2\mathbb{1}_{\mathcal{X}_{\text{on}}}]}}\sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H}H^2 \vee \sum_{i=1}^{t-1}\mathbb{E}_{d_h^{(i)}}[(\delta_h^{(t)})^2\mathbb{1}_{\mathcal{X}_{\text{on}}}]} \\
&\leq \sqrt{H\mathsf{SEC}(\mathcal{F}_{\text{on}}, \Pi, T)}\sqrt{\beta H^3 T}
\end{aligned}
$$

$$\leq \sqrt{\beta H^4 T \mathsf{SEC}(\mathcal{F}_{\text{on}}, \Pi, T)}. \tag{82}$$

Within the third-last inequality, the first term can be bounded by H times the SEC of Xie et al. (2022), almost by definition of the SEC. The second term is bounded by Lemma 14.

The offline term is bounded by the offline data. We perform a similar Cauchy-Schwarz and change of measure argument as Tan and Xu (2024) to see that:

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{d_h^{(t)}}[\delta_h^{(t)} \mathbb{1}_{\mathcal{X}_{\text{off}}}] &= \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{d_h^{(t)}}[\delta_h^{(t)} \mathbb{1}_{\mathcal{X}_{\text{off}}}] \left(\frac{N_{\text{off}} \mathbb{E}_{\mu}[(\delta_h^{(t)})^2 \mathbb{1}_{\mathcal{X}_{\text{off}}}] + \sum_{i=1}^{t-1} \mathbb{E}_{d_h^{(i)}}[(\delta_h^{(t)})^2 \mathbb{1}_{\mathcal{X}_{\text{off}}}]}{N_{\text{off}} \mathbb{E}_{\mu}[(\delta_h^{(t)})^2 \mathbb{1}_{\mathcal{X}_{\text{off}}}] + \sum_{i=1}^{t-1} \mathbb{E}_{d_h^{(i)}}[(\delta_h^{(t)})^2 \mathbb{1}_{\mathcal{X}_{\text{off}}}]} \right)^{1/2} \\
&\leq \sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H} \frac{\mathbb{E}_{d_h^{(i)}}[\delta_h^{(t)} \mathbb{1}_{\mathcal{X}_{\text{off}}}]^2}{N_{\text{off}} \mathbb{E}_{\mu}[(\delta_h^{(t)})^2 \mathbb{1}_{\mathcal{X}_{\text{off}}}] + \sum_{i=1}^{t-1} \mathbb{E}_{d_h^{(i)}}[(\delta_h^{(t)})^2 \mathbb{1}_{\mathcal{X}_{\text{off}}}]}} \\
&\quad \cdot \sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H} \left(N_{\text{off}} \mathbb{E}_{\mu}[(\delta_h^{(t)})^2 \mathbb{1}_{\mathcal{X}_{\text{off}}}] + \sum_{i=1}^{t-1} \mathbb{E}_{d_h^{(i)}}[(\delta_h^{(t)})^2 \mathbb{1}_{\mathcal{X}_{\text{off}}}] \right)} \\
&\leq \sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H} \frac{\mathbb{E}_{d_h^{(i)}}[\delta_h^{(t)} \mathbb{1}_{\mathcal{X}_{\text{off}}}]^2}{N_{\text{off}} \mathbb{E}_{\mu}[(\delta_h^{(t)})^2 \mathbb{1}_{\mathcal{X}_{\text{off}}}]}} \\
&\quad \cdot \sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H} \left(N_{\text{off}} \mathbb{E}_{\mu}[(\delta_h^{(t)})^2 \mathbb{1}_{\mathcal{X}_{\text{off}}}] + \sum_{i=1}^{t-1} \mathbb{E}_{d_h^{(i)}}[(\delta_h^{(t)})^2 \mathbb{1}_{\mathcal{X}_{\text{off}}}] \right)} \\
&\leq \sqrt{c_{\text{off}}(\mathcal{F} \mathbb{1}_{\mathcal{X}_{\text{off}}}) HT / N_{\text{off}}} \sqrt{\beta H^3 T} \\
&\leq \sqrt{\beta H^4 c_{\text{off}}(\mathcal{F} \mathbb{1}_{\mathcal{X}_{\text{off}}}) T^2 / N_{\text{off}}}.
\end{aligned} \tag{83}$$

Note that the first term of the penultimate line follows directly from Definition 5 on the partial all-policy concentrability, and the second term term follows directly from Lemma 16.

D.4. Auxiliary lemmas

We also require the following helper lemma:

Lemma 15 (Optimism and in-sample error control for critics targeting $Q^{\pi^{(t)}}$ in hybrid RL)**.** *With probability at least $1-\delta$, for all $t \in [T]$, we have that for all $h = 1, \ldots, H$, a critic targeting $Q^{\pi^{(t)}}$ in the same way as defined in Algorithm 3 achieves*

$$N_{\text{off}} \mathbb{E}_{\mu_h}\left[\left(f_h^{(t)} - \mathcal{T}_h f_{h+1}^{(t)} \right)^2 \right] + \sum_{i=1}^{t-1} \mathbb{E}_{d^{\pi^{(i)}}}\left[\left(f_h^{(t)} - \mathcal{T}_h f_{h+1}^{(t)} \right)^2 \right] \leq O\left(H^2 \beta\right)$$

by choosing $\beta = c_1 \left(\log[NH\mathcal{N}_{\mathcal{F},(\mathcal{T}^{\Pi})^T \mathcal{F}}(\rho)/\delta] + N\rho \right)$ for some constant c_1, where $N = N_{\text{off}} + T$.

Proof. The proof is analogous to that of Lemma 1 in Tan and Xu (2024). We apply property (ii) of Lemma 7 in the following way: we append a sequence of functions generated from offline samples $1, ..., N_{\text{off}}$ to the start of the sequence of T online samples.

Let $f^{(t)}$ be a sequence of critics in $\mathcal{F}$, defined as follows. Arrange the offline samples in any order. For each $n \in [N_{\text{off}}]$, define $f^{(n)}$ to be any function in the confidence sets constructed by the first n offline episodes.

Now for each $t = N_{\text{off}} + 1, ..., N$, define $f^{(t)} := f^{(t - N_{\text{off}})} \in \mathcal{F}^{(t)}$. As Lemma 7 shows that property (ii) holds for all $\tau \in [N]$, it must also hold for all $\tau = N_{\text{off}} + 1, \ldots, N$. □

Lemma 16 (Optimism and in-sample error control for critics targeting Q^* in hybrid RL)**.** *With probability at least $1-\delta$, for all $t \in [T]$, we have that for all $h = 1, \ldots, H$, a critic targeting Q^* in the same way as defined in Algorithm 5 achieves*

(i) $\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} \in \mathcal{F}_h^{(t)}$, and $f_h^{(t)} \geq \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}$,

$$\textit{(ii)}\ N_{\text{off}}\mathbb{E}_{\mu_h}\left[\left(f_h^{(t)}-\mathcal{T}_h f_{h+1}^{(t)}\right)^2\right]+\sum_{i=1}^{t-1}\mathbb{E}_{d^{\pi^{(i)}}}\left[\left(f_h^{(t)}-\mathcal{T}_h f_{h+1}^{(t)}\right)^2\right]\le O\left(H^2\beta\right),$$

by choosing $\beta = c_1\left(\log\left[NH\mathcal{N}_{\mathcal{F},\mathcal{TF}}(\rho)/\delta\right]+N\rho\right)$ *for some constant* c_1*, where* $N = N_{\text{off}}+T$*. If the critic is non-optimistic as in Algorithm 4, only (ii) holds.*

Proof. The proof is analogous to that of Lemma 1 in Tan and Xu (2024). We apply Lemma 14 in the following way: we append a sequence of functions generated from offline samples $1, ..., N_{\text{off}}$ to the start of the sequence of T online samples.

Let $f^{(t)}$ be a sequence of critics in $\mathcal{F}$, defined as follows. Arrange the offline samples in any order. For each $n \in [N_{\text{off}}]$, define $f^{(n)}$ to be any function in the confidence sets constructed by the first n offline episodes.

Now for each $t = N_{\text{off}}+1, ..., N$, define $f^{(t)} := f^{(t-N_{\text{off}})} \in \mathcal{F}^{(t)}$. As Lemma 14 shows that (i) and (ii) hold for all $\tau \in [N]$, they must also hold for all $\tau = N_{\text{off}}+1, \dots, N$. □

E. Concentration of the Empirical Loss

Lemma 17 (Modified Lemma G.1, Xiong et al. (2023))**.** *For any* $f \in \mathcal{F}$*,* $\pi \in \Pi$*,* $h \in [H]$ *and* $t_1 \le t_2 \in [T]$*, we have with probability* $1-\delta$ *that*

$$0 \le \mathcal{L}_h^{(t_1:t_2,\pi)}(\mathcal{T}_h^\pi f_{h+1}, f_{h+1}) - \min_{f_h'\in\mathcal{F}_h}\mathcal{L}_h^{(t_1:t_2,\pi)}(f_h', f_{h+1}) \le H^2\beta,$$

where $\beta = c_1\left(\log\left[NH\mathcal{N}_{\mathcal{F},\mathcal{G}}(\rho)/\delta\right]+N\rho\right)$ *for some constant* c_1*. Note that* $\mathcal{G} = \mathcal{TF}$ *if* π *is the greedy policy with respect to* f*, and* $\mathcal{G} = (\mathcal{T}^\Pi)^T\mathcal{F}$ *otherwise.*

Proof. The proof is for the most part similar to the proof of Lemma G.1 for Xiong et al. (2023), which is in turn an analogue of Lemma 40 in Jin et al. (2021). However, we take into account the fact that the Bellman operator we are concerned is the Bellman operator for policy π. That is, we are concerned with $\mathcal{T}^\pi$, not $\mathcal{T}$. For completeness and thoroughness, we provide the full proof below.

Write $\ell_h^{(i)}(f,g,\pi) = g_h(s_h^{(i)},a_h^{(i)}) - r_h(s_h^{(i)},a_h^{(i)}) - f_{h+1}(s_{h+1}^{(i)},\pi_{h+1}(s_{h+1}^{(i)}))$ for the TD error at timestep h and trajectory i for policy π, and $\delta_h^{(f,g,\pi)}(s_h^{(i)},a_h^{(i)}) = \mathbb{E}_{s_{h+1}^{(i)}\sim d^{\pi^{(i)}}}[\ell_h^{(i)}(f,g,\pi)]$ for its expectation over $s_{h+1}^{(i)} \sim d^{\pi^{(i)}}$, known as the Bellman error between f and g.

Now write $X_h^{(i)}(f,g,\pi) = \ell_h^{(i)}(f,g,\pi)^2 - \ell_h^{(i)}(f,\mathcal{T}^\pi f,\pi)^2$ for the difference in the squared TD error between f_h and g_h and the squared TD error between $\mathcal{T}_h^\pi f_{h+1}$ and f_h. Note that this is bounded by $O(H^2)$. We can then show that

$$\begin{aligned}
&\mathbb{E}_{s_{h+1}^{(i)}\sim d^{\pi^{(i)}}}[X_h^{(i)}(f,g,\pi)]\\
&= \mathbb{E}_{s_{h+1}^{(i)}\sim d^{\pi^{(i)}}}[(\ell_h^{(i)}(f,g,\pi) - \ell_h^{(i)}(f,\mathcal{T}_h^\pi f,\pi))\cdot(\ell_h^{(i)}(f,g,\pi) + \ell_h^{(i)}(f,\mathcal{T}_h^\pi f,\pi))]\\
&= \mathbb{E}_{s_{h+1}^{(i)}\sim d^{\pi^{(i)}}}[\mathbb{E}_{s_{h+1}^{(i)}\sim d^{\pi^{(i)}}}[\ell_h^{(i)}(f,g,\pi)]\cdot\ell_h^{(i)}(f,g,\pi)]\\
&= \delta_h^{(f,g,\pi)}(s_h^{(i)},a_h^{(i)})^2,
\end{aligned} \tag{84}$$

where the second-last equality follows from noting that (Chen et al., 2022)

$$\begin{aligned}
\ell_h^{(i)}(f,g,\pi) - \mathbb{E}_{s_{h+1}^{(i)}\sim d^{\pi^{(i)}}}[\ell_h^{(i)}(f,g,\pi)] &= g_h(s_h^{(i)},a_h^{(i)}) - r_h(s_h^{(i)},a_h^{(i)}) - f_{h+1}(s_{h+1}^{(i)},\pi_{h+1}(s_{h+1}^{(i)}))\\
&\quad - g_h(s_h^{(i)},a_h^{(i)}) + \mathcal{T}_h^\pi f_{h+1}(s_h^{(i)},a_h^{(i)})\\
&= \mathcal{T}_h^\pi f_{h+1}(s_h^{(i)},a_h^{(i)}) - r_h(s_h^{(i)},a_h^{(i)}) - f_{h+1}(s_{h+1}^{(i)},\pi_{h+1}(s_{h+1}^{(i)}))\\
&= \ell_h^{(i)}(f,\mathcal{T}_h^\pi f,\pi).
\end{aligned} \tag{85}$$

We use this property again in the fourth equality below to show that

$$\mathbb{E}_{s_{h+1}^{(i)}\sim d^{\pi^{(i)}}}[X_h^{(i)}(f,g,\pi)^2] = \mathbb{E}_{s_{h+1}^{(i)}\sim d^{\pi^{(i)}}}[(\ell_h^{(i)}(f,g,\pi) - \ell_h^{(i)}(f,\mathcal{T}_h^\pi f,\pi))^2\cdot(\ell_h^{(i)}(f,g,\pi) + \ell_h^{(i)}(f,\mathcal{T}_h^\pi f,\pi))^2]$$

$$\leq 4H^2\mathbb{E}_{s_{h+1}^{(i)}\sim d^{\pi^{(i)}}}[(\ell_h^{(i)}(f,g,\pi)-\ell_h^{(i)}(f,\mathcal{T}_h^\pi f,\pi))^2]$$
$$= 4H^2\delta_h^{(f,g,\pi)}(s_h^{(i)},a_h^{(i)})^2$$
$$= \mathbb{E}_{s_{h+1}^{(i)}\sim d^{\pi^{(i)}}}[X_h^{(i)}(f,g,\pi)]. \tag{86}$$

We can then apply Freedman's inequality (Jin et al., 2021; Chen et al., 2022) and a union bound over the value function class $\mathcal{V}$ to see that for any $f\in\mathcal{F}$ and $g\in\mathcal{G}$,

$$\left|\sum_{i=t_1}^{t_2} X_h^{(i)}(f,g,\pi)-\sum_{i=t_1}^{t_2}\mathbb{E}_{s_{h+1}^{(i)}\sim d^{\pi^{(i)}}}[X_h^{(i)}(f,g,\pi)]\right| \lesssim H\sqrt{\iota\sum_{i=t_1}^{t_2}\mathbb{E}_{s_{h+1}^{(i)}\sim d^{\pi^{(i)}}}[X_h^{(i)}(f,g,\pi)]}+H^2\iota \tag{87}$$

holds with probability at least $1-\delta$, where $\iota=\log(HT^2\mathcal{N}_{\mathcal{F},\mathcal{G}}(1/T)/\delta)$.

The result now holds by observing that

$$\mathbb{E}_{s_{h+1}^{(i)}\sim d^{\pi^{(i)}}}[X_h^{(i)}(f,g,\pi)]\geq 0$$

and that

$$\sum_{i=t_1}^{t_2} X_h^{(i)}(f,g,\pi)\leq O(H^2\iota+H)\leq H^2\beta$$

for any $f\in\mathcal{F}$ and $g\in\mathcal{G}$ with probability at least $1-\delta$.

□

Lemma 18 (Modified Lemma G.2, Xiong et al. (2023)). *Let $t_1\leq t_2\in[T]$, $\pi\in\Pi$, and $f\in\mathcal{F}$. If it holds with probability at least $1-2\delta$ that*

$$\mathcal{L}_h^{(t_1:t_2,\pi)}(f_h,f_{h+1})-\mathcal{L}_h^{(t_1:t_2,\pi)}(\mathcal{T}_h^\pi f_{h+1},f_{h+1})\leq CH^2\beta,$$

then it also holds with probability at least $1-2\delta$ that:

$$\sum_{i=t_1}^{t_2}\delta_h^{(f,f,\pi)}(s_h^{(i)},a_h^{(i)})^2\leq C_1H^2\beta,\qquad \sum_{i=t_1}^{t_2}\mathbb{E}_{d_h^{\pi^{(i)}}}\left[\left(\delta_h^{(f,f,\pi)}\right)^2\right]\leq C_1H^2\beta,$$

where $C_1=C+1$ if $C\in[1,100]$, or $2C$ for all constants $C\geq 2$, and $\beta=c_1(\log[NH\mathcal{N}_{\mathcal{F},\mathcal{G}}(\rho)/\delta]+N\rho)$ for some constant c_1. Note that $\mathcal{G}=\mathcal{T}\mathcal{F}$ if π is the greedy policy with respect to f, and $\mathcal{G}=(\mathcal{T}^\Pi)^T\mathcal{F}$ otherwise.

Proof. The proof is similar to that of Lemma G.2 in Xiong et al. (2023). By the same argument as in Lemma 17, we apply Freedman's inequality (Jin et al., 2021; Chen et al., 2022) and a union bound over a $1/T$-net (or $1/N$ net in the hybrid case) of $(\mathcal{F},\mathcal{G})$ to see that for any $f\in\mathcal{F}$ and $g\in\mathcal{G}$,

$$\left|\sum_{i=t_1}^{t_2} X_h^{(i)}(f,g,\pi)-\sum_{i=t_1}^{t_2}\mathbb{E}_{s_{h+1}^{(i)}\sim d^{\pi^{(i)}}}[X_h^{(i)}(f,g,\pi)]\right| \lesssim H\sqrt{\iota\sum_{i=t_1}^{t_2}\mathbb{E}_{s_{h+1}^{(i)}\sim d^{\pi^{(i)}}}[X_h^{(i)}(f,g,\pi)]}+H^2\iota \tag{88}$$

holds with probability at least $1-\delta$.

By assumption, we have that

$$\sum_{i=t_1}^{t_2} X_h^{(i)}(f,f,\pi)=\sum_{i=t_1}^{t_2}\left(\ell_h^{(i)}(f,f,\pi)^2-\ell_h^{(i)}(f,\mathcal{T}_h^\pi f,\pi)^2\right)\leq CH^2\beta. \tag{89}$$

If $C\in[1,100]$ and therefore is a relatively small bounded constant, we can choose c_1 large enough in the definition of β so that there is some $\widetilde{f},\widetilde{g}$ in the covering of $(\mathcal{F},\mathcal{G})$ such that

$$\left|\sum_{i=t_1}^{t_2} X_h^{(i)}(f,f,\pi)-\sum_{i=t_1}^{t_2} X_h^{(i)}(\widetilde{f},\widetilde{g},\pi)\right|\leq O(H),\text{ and }\sum_{i=t_1}^{t_2} X_h^{(i)}(\widetilde{f},\widetilde{g},\pi)\leq CH^2\beta+O(H), \tag{90}$$

and consequently that

$$\sum_{i=t_1}^{t_2} X_h^{(i)}(\widetilde{f},\widetilde{g},\pi) \le (C+1/2)H^2\beta, \text{ and } \sum_{i=t_1}^{t_2} \mathbb{E}_{s_{h+1}^{(i)}\sim d^{\pi^{(i)}}}[X_h^{(i)}(f,f,\pi)] \le (C+1/2)H^2\beta + O(H) \le (C+1)H^2\beta.$$

As in Xiong et al. (2023) and Jin et al. (2021), the argument for the case where $C > 100$ follows by accepting a 2-approximation where $C_1 = 2C$, and the argument for $\sum_{i=1}^{t-1} \mathbb{E}_{d_h^{\pi^{(i)}}}\left[\left(\delta_h^{(t)}\right)^2\right]$ also follows analogously by taking expectations.

Note that taking $t_1 = 1, t_2 = t-1, \pi = \pi^{(t-1)}, f = f^{(t)}$ or $t_1 = 1, t_2 = t-1, \pi = \pi^f, f = f^{(t)}$ gives results of direct importance to us. □

Lemma 19 (Modified Lemma G.3, Xiong et al. (2023))**.** *Let $t_1 \le t_2 \in [T]$, $\pi \in \Pi$, and $f \in \mathcal{F}$. If it holds with probability at least $1-2\delta$ that*

$$\mathcal{L}_h^{(t_1:t_2,\pi)}(f_h, f_{h+1}) - \mathcal{L}_h^{(t_1:t_2,\pi)}(\mathcal{T}_h^\pi f_{h+1}, f_{h+1}) \ge CH^2\beta,$$

then it also holds with probability at least $1-2\delta$ that:

$$\sum_{i=t_1}^{t_2} \delta_h^{(f,f,\pi)}(s_h^{(i)}, a_h^{(i)})^2 \ge C_1H^2\beta, \quad \sum_{i=t_1}^{t_2} \mathbb{E}_{d_h^{\pi^{(i)}}}\left[\left(\delta_h^{(f,f,\pi)}\right)^2\right] \ge C_1H^2\beta,$$

where $C_1 = C+1$ if $C \in [1,100]$, or $2C$ for all constants $C \ge 2$. , and $\beta = c_1(\log[NH\mathcal{N}_{\mathcal{F},\mathcal{G}}(\rho)/\delta] + N\rho)$ for some constant c_1. Note that $\mathcal{G} = \mathcal{T}\mathcal{F}$ if π is the greedy policy with respect to f, and $\mathcal{G} = (\mathcal{T}^\Pi)^T\mathcal{F}$ otherwise.

Proof. The proof is similar to that of Lemma G.3 in Xiong et al. (2023). By the same argument as in Lemma 17, we apply Freedman's inequality (Jin et al., 2021; Chen et al., 2022) and a union bound over a $1/T$-net (or $1/N$ net in the hybrid case) of $(\mathcal{F},\mathcal{G})$ to see that for any $f \in \mathcal{F}$ and $g \in \mathcal{T}^\pi\mathcal{F}$,

$$\left|\sum_{i=t_1}^{t_2} X_h^{(i)}(f,g,\pi) - \sum_{i=t_1}^{t_2} \mathbb{E}_{s_{h+1}^{(i)}\sim d^{\pi^{(i)}}}[X_h^{(i)}(f,g,\pi)]\right| \lesssim H\sqrt{\iota\sum_{i=t_1}^{t_2} \mathbb{E}_{s_{h+1}^{(i)}\sim d^{\pi^{(i)}}}[X_h^{(i)}(f,g,\pi)]} + H^2\iota \tag{91}$$

holds with probability at least $1-\delta$.

By assumption, we have that

$$\sum_{i=t_1}^{t_2} X_h^{(i)}(f,f,\pi) = \sum_{i=t_1}^{t_2}\left(\ell_h^{(i)}(f,f,\pi)^2 - \ell_h^{(i)}(f,\mathcal{T}_h^\pi f,\pi)^2\right) \ge CH^2\beta.$$

If $C \in [1,100]$ and therefore is a relatively small bounded constant, we can choose c_1 large enough in the definition of β so that there is some $\widetilde{f},\widetilde{g}$ in the covering of $(\mathcal{F},\mathcal{G})$ such that

$$\left|\sum_{i=t_1}^{t_2} X_h^{(i)}(f,f,\pi) - \sum_{i=t_1}^{t_2} X_h^{(i)}(\widetilde{f},\widetilde{g},\pi)\right| \le O(H), \text{ and } \sum_{i=t_1}^{t_2} X_h^{(i)}(\widetilde{f},\widetilde{g},\pi) \ge CH^2\beta - O(H), \tag{92}$$

and consequently that

$$\sum_{i=t_1}^{t_2} X_h^{(i)}(\widetilde{f},\widetilde{g},\pi) \ge (C-1/2)H^2\beta, \text{ and } \sum_{i=t_1}^{t_2} \mathbb{E}_{s_{h+1}^{(i)}\sim d^{\pi^{(i)}}}[X_h^{(i)}(f,f,\pi)] \ge (C-1/2)\beta - O(H) \ge (C-1)H^2\beta. \tag{93}$$

As in Xiong et al. (2023) and Jin et al. (2021), the argument for the case where $C > 100$ follows by accepting a 2-approximation where $C_1 = C/2$, and the argument for $\sum_{i=1}^{t-1} \mathbb{E}_{d_h^{\pi^{(i)}}}\left[\left(\delta_h^{(t)}\right)^2\right]$ also follows analogously by taking expectations.

Note that taking $t_1 = 1, t_2 = t-1, \pi = \pi^{(t-1)}, f = f^{(t)}$ or $t_1 = 1, t_2 = t-1, \pi = \pi^f, f = f^{(t)}$ gives results of direct importance to us. □

F. What If We Target $Q^{\pi^{(t)}}$ Instead of Q^*?

Algorithm 6 NORA-π

Input: Offline dataset $\mathcal{D}_{\text{off}}$, samples sizes T, N_{off}, function class $\mathcal{F}$ and confidence width $\beta > 0$
Initialize: $\mathcal{F}^{(0)} \leftarrow \mathcal{F}$, $\mathcal{D}_h^{(0)} \leftarrow \emptyset, \forall h \in [H]$, $\eta = \Theta(\sqrt{\log|\mathcal{A}|H^{-2}T^{-1}})$, $\pi^{(1)} \propto 1$.
for episode $t = 1, 2, \ldots, T$ **do**
  Select critic $f_h^{(t)}(s,a) := \operatorname{argmax}_{f_h \in \mathcal{F}_h^{(t_{\text{last}})}} f_h(s,a)$ for all s, a.
  Play policy $\pi^{(t)}$ for one episode and obtain trajectory $(s_1^{(t)}, a_1^{(t)}, r_1^{(t)}), \ldots, (s_H^{(t)}, a_H^{(t)}, r_H^{(t)})$.
  Update dataset $\mathcal{D}_h^{(t)} \leftarrow \mathcal{D}_h^{(t-1)} \cup \{(s_h^{(t)}, a_h^{(t)}, r_h^{(t)}, s_{h+1}^{(t)})\}, \forall h \in [H]$.
  if there exists some h such that $\mathcal{L}_h^{(t,\pi^{(t)})}(f_h^{(t)}, f_{h+1}^{(t)}) - \min_{f'_h \in \mathcal{F}_h} \mathcal{L}_h^{(t,\pi^{(t)})}(f'_h, f_{h+1}^{(t)}) \geq 5H^2\beta$ **then**
    Compute confidence set for Bellman operator $\mathcal{T}^{\pi^{(t)}}$, set $\mathcal{F}^{(t)} \leftarrow \mathcal{F}^{(t,\pi^{(t)})}$.

$$\mathcal{F}^{(t,\pi^{(t)})} \leftarrow \left\{ f \in \mathcal{F} : \mathcal{L}_h^{(t,\pi^{(t)})}(f_h, f_{h+1}) - \min_{f'_h \in \mathcal{F}_h} \mathcal{L}_h^{(t,\pi^{(t)})}(f'_h, f_{h+1}) \leq H^2\beta \quad \forall h \in [H] \right\},$$

$$\text{where } \mathcal{L}_h^{(t,\pi^{(t)})}(f, f') := \sum_{(s,a,r,s') \in \mathcal{D}_h^{(t)} \cup \mathcal{D}_{\text{off},h}} \left(f(s,a) - r - f'(s', \pi_{h+1}^{(t)}(s')) \right)^2, \forall f \in \mathcal{F}_h, f' \in \mathcal{F}_{h+1}.$$

    Set $t_{\text{last}} := t$, increment number of updates $N_{\text{updates}}^{(t)} := N_{\text{updates}}^{(t-1)} + 1$.
  else
    Set $N_{\text{updates}}^{(t)} := N_{\text{updates}}^{(t-1)}$, $\mathcal{F}^{(t)} := \mathcal{F}^{(t-1)}$.
  end if
  Select policy $\pi_h^{(t+1)} \propto \pi_h^{(t)} \exp(\eta f_h^{(t)})$.
end for

F.1. But can we bound the number of critic updates?

We attempt to show a similar result to Lemma 13 for the Q-function confidence set $\mathcal{F}^{(t,\pi^{(t)})}$ targeting $\pi^{(t)}$. However, we will later see that we run into an issue.

Fix some $h \in [H]$ for now. For simplicity, write $K_h = N_{\text{updates},h}(T)$ for the total number of updates induced by updating the Q-function class targeting $\pi^{(t)}$, and $t_{1,h}, ..., t_{K_h,h}$ the update times for $f_h^{(t)}$, with $t_{0,h} = 0$. By definition, at every $t_{k,h}$, we have

$$\mathcal{L}_h^{(t_{k,h},\pi^{(t_{k,h})})}\left(f_h^{(t_{k,h})}, f_{h+1}^{(t_{k,h})}\right) - \min_{f'_h \in \mathcal{F}_h} \mathcal{L}_h^{(t_{k,h},\pi^{(t_{k,h})})}\left(f'_h, f_{h+1}^{(t_{k,h})}\right) \geq 5H^2\beta \tag{94}$$

An application of Lemma 17 yields

$$0 \leq \mathcal{L}_h^{\left(t_{k,h},\pi^{(t_{k,h})}\right)}\left(\mathcal{T}_h^{\pi^{(t_{k,h})}} f_{h+1}^{(t_{k,h})}, f_{h+1}^{(t_{k,h})}\right) - \min_{f'_h \in \mathcal{F}_h} \mathcal{L}_h^{\left(t_{k,h},\pi^{(t_{k,h})}\right)}\left(f'_h, f_{h+1}^{(t_{k,h})}\right) \leq H^2\beta,$$

$$\min_{f'_h \in \mathcal{F}_h} \mathcal{L}_h^{\left(t_{k,h},\pi^{(t_{k,h})}\right)}\left(f'_h, f_{h+1}^{(t_{k,h})}\right) \geq \mathcal{L}_h^{\left(t_{k,h},\pi^{(t_{k,h})}\right)}\left(\mathcal{T}_h^{\pi^{(\pi^{(t_{k,h})})}} f_{h+1}^{(t_{k,h})}, f_{h+1}^{(t_{k,h})}\right) - H^2\beta,$$

$$\mathcal{L}_h^{(t_{k,h},\pi^{(t_{k,h})})}\left(f_h^{(t_{k,h})}, f_{h+1}^{(t_{k,h})}\right) - \mathcal{L}_h^{\left(t_{k,h},\pi^{(t_{k,h})}\right)}\left(\mathcal{T}_h^{\pi^{(\pi^{(t_{k,h})})}} f_{h+1}^{(t_{k,h})}, f_{h+1}^{(t_{k,h})}\right) \geq 4H^2\beta.$$

From the above, we can now establish that

$$\begin{aligned}
&\mathcal{L}_h^{(t_{k-1,h}+1:t_{k,h},\pi^{(t_{k,h})})}\left(f_h^{(t_{k,h})}, f_{h+1}^{(t_{k,h})}\right) - \mathcal{L}_h^{\left(t_{k-1,h}+1:t_{k,h},\pi^{(t_{k,h})}\right)}\left(\mathcal{T}_h^{\pi^{(t_{k,h})}} f_{h+1}^{(t_{k,h})}, f_{h+1}^{(t_{k,h})}\right) \\
&= \mathcal{L}_h^{(t_{k-1,h}+1:t_{k,h},\pi^{(t_{k,h})})}\left(f_h^{(t_{k-1,h}+1)}, f_{h+1}^{(t_{k-1,h}+1)}\right) - \mathcal{L}_h^{\left(t_{k-1,h}+1:t_{k,h},\pi^{(t_{k,h})}\right)}\left(\mathcal{T}_h^{\pi^{(t_{k,h})}} f_{h+1}^{(t_{k-1,h}+1)}, f_{h+1}^{(t_{k-1,h}+1)}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \mathcal{L}_h^{(t_{k,h},\pi^{(t_{k,h})})}\left(f_h^{(t_{k-1,h}+1)}, f_{h+1}^{(t_{k-1,h}+1)}\right) - \mathcal{L}_h^{\left(t_{k,h},\pi^{(t_{k,h})}\right)}\left(\mathcal{T}_h^{\pi^{(t_{k,h})}} f_{h+1}^{(t_{k-1,h}+1)}, f_{h+1}^{(t_{k-1,h}+1)}\right) \\
&\quad - \left(\mathcal{L}_h^{(t_{k-1,h},\pi^{(t_{k,h})})}\left(f_h^{(t_{k-1,h}+1)}, f_{h+1}^{(t_{k-1,h}+1)}\right) + \mathcal{L}_h^{(t_{k-1,h},\pi^{(t_{k,h})})}\left(\mathcal{T}_h^{\pi^{(t_{k,h})}} f_{h+1}^{(t_{k-1,h}+1)}, f_{h+1}^{(t_{k-1,h}+1)}\right)\right) \\
&= \mathcal{L}_h^{(t_{k,h},\pi^{(t_{k,h})})}\left(f_h^{(t_{k,h})}, f_{h+1}^{(t_{k,h})}\right) - \mathcal{L}_h^{\left(t_{k,h},\pi^{(t_{k,h})}\right)}\left(\mathcal{T}_h^{\pi^{(t_{k,h})}} f_{h+1}^{(t_{k,h})}, f_{h+1}^{(t_{k,h})}\right) \\
&\quad - \left(\mathcal{L}_h^{(t_{k-1,h},\pi^{(t_{k,h})})}\left(f_h^{(t_{k-1,h}+1)}, f_{h+1}^{(t_{k-1,h}+1)}\right) + \mathcal{L}_h^{(t_{k-1,h},\pi^{(t_{k,h})})}\left(\mathcal{T}_h^{\pi^{(t_{k,h})}} f_{h+1}^{(t_{k-1,h}+1)}, f_{h+1}^{(t_{k-1,h}+1)}\right)\right) \\
&\geq 4H^2\beta - H^2\beta = 3H^2\beta
\end{aligned} \tag{95}
$$

We substitute $f_h^{(t_{k,h})}$ for $f_h^{(t_{k-1,h}+1)}$ in the second equality, and obtain the last inequality via the inequality established in the previous argument and an application of Lemma 17 on $t_2 = t_{k-1,h}, \pi = \pi^{(t_{k,h})}, f = f^{(t_{k-1,h}+1)}$.

Therefore, for any t such that $t_{k-1,h} < t \leq t_{k,h}$, this argument and noting that $f_h^{(t_{k-1,h}+1)} = ... = f_h^{(t)} = ... = f_h^{(t_{k,h})}$ yields

$$
\mathcal{L}_h^{(t_{k-1,h}+1:t_{k,h},\pi^{(t_{k,h})})}\left(f_h^{(t)}, f_{h+1}^{(t)}\right) - \mathcal{L}_h^{\left(t_{k-1,h}+1:t_{k,h},\pi^{(t_{k,h})}\right)}\left(\mathcal{T}_h^{\pi^{(t_{k,h})}} f_{h+1}^{(t)}, f_{h+1}^{(t)}\right) \geq 3H^2\beta. \tag{96}
$$

An application of Lemma 19 while noting that $f_h^{(t_{k-1,h}+1)} = ... = f_h^{(t)} = ... = f_h^{(t_{k,h})}$ yields

$$
\sum_{i=t_{k-1,h}+1}^{t_{k,h}} \left(f_h^{(i)} - \mathcal{T}_h^{\pi^{(t_{k,h})}} f_{h+1}^{(i)}\right)^2 (s_h^{(i)}, a_h^{(i)}) = \sum_{i=t_{k-1,h}+1}^{t_{k,h}} \left(f_h^{(t_{k,h})} - \mathcal{T}_h^{\pi^{(t_{k,h})}} f_{h+1}^{(t_{k,h})}\right)^2 (s_h^{(i)}, a_h^{(i)}) \geq 2H^2\beta. \tag{97}
$$

Summing over all $t_{1,h}, ..., t_{K,h}$ yields

$$
\sum_{t=1}^{T} \left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t_{\text{next}})}} f_{h+1}^{(t)}\right)^2 (s_h^{(t)}, a_h^{(t)}) = \sum_{k=1}^{K_h} \sum_{i=t_{k-1,h}+1}^{t_{k,h}} \left(f_h^{(i)} - \mathcal{T}_h^{\pi^{(t_{k,h})}} f_{h+1}^{(i)}\right)^2 (s_h^{(i)}, a_h^{(i)}) \geq 2(K_h - 1)H^2\beta. \tag{98}
$$

By Lemma 14, we have that

$$
\sum_{i=1}^{t-1} \left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t-1)}} f_{h+1}^{(t)}\right)^2 (s_h^{(i)}, a_h^{(i)}) \leq O(H^2\beta). \tag{99}
$$

Invoking the squared distributional Bellman eluder dimension definition yields

$$
\sum_{t=1}^{T} \left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t-1)}} f_{h+1}^{(t)}\right)^2 (s_h^{(t)}, a_h^{(t)}) \leq O(dH^2\beta \log T). \tag{100}
$$

So we have established that

$$
2(K_h - 1)H^2\beta \leq \sum_{t=1}^{T} \left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t_{\text{next}})}} f_{h+1}^{(t)}\right)^2 (s_h^{(t)}, a_h^{(t)}),
$$

$$
\text{and } \sum_{t=1}^{T} \left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t-1)}} f_{h+1}^{(t)}\right)^2 (s_h^{(t)}, a_h^{(t)}) \leq O(dH^2 \log T).
$$

However, it remains unclear how one can relate

$$
\sum_{t=1}^{T} \left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t_{\text{next}})}} f_{h+1}^{(t)}\right)^2 (s_h^{(t)}, a_h^{(t)})
$$

$$\text{to } \sum_{t=1}^{T} \left(f_h^{(t)} - \mathcal{T}_h^{\pi^{(t-1)}} f_{h+1}^{(t)} \right)^2 (s_h^{(t)}, a_h^{(t)}).$$

We would like the former to be no greater than the latter, but that does not necessarily hold, as $\pi^{(t-1)}$ is closer to the target by which $\mathcal{F}^{(t)}$ was constructed, $\pi^{(t_{\text{last}})}$, than $\pi^{(t_{\text{next}})}$. So if anything, it is likely that the Bellman error under the Bellman operator for $\pi^{(t_{\text{next}})}$ is greater than that for $\pi^{(t-1)}$. It is therefore difficult to say anything with regard to the number of updates for each h.

F.2. But can we control the negative Bellman error?

There is another obstacle. It is unclear how to control the negative Bellman error under the occupancy measure of the optimal policy, given the far more limited form of optimism in 7. This is because the more limited form of optimism only allows us to show:

$$-\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[f_h^{(t)} - \mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)} \right] \le \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\left\langle f_{h+1}^{(t)}(s', \cdot), \pi_{h+1}^{(t)}(\cdot|s') - \pi_{h+1}^{(t_{\text{last}})}(\cdot|s') \right\rangle \right]. \tag{101}$$

To see this, observe that by Lemma 7, $\mathcal{T}_h^{\pi^{(t_{\text{last}})}} f_{h+1}^{(t)}(s,a) \le f_h^{(t)}(s,a)$. Therefore,

$$\begin{aligned}
-\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[f_h^{(t)} - \mathcal{T}^{\pi^{(t)}} f_{h+1}^{(t)} \right] &= \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[r_h(s,a) + f_{h+1}^{(t)}(s', \pi_{h+1}^{(t)}(s')) - f_h^{(t)}(s,a) \right] \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}(s,a) - f_h^{(t)}(s,a) \right] \\
&\le \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\mathcal{T}_h^{\pi^{(t)}} f_{h+1}^{(t)}(s,a) - \mathcal{T}_h^{\pi^{(t_{\text{last}})}} f_{h+1}^{(t)}(s,a) \right] \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[r_h(s,a) + f_{h+1}^{(t)}(s', \pi_{h+1}^{(t)}(s')) - r_h(s,a) - f_{h+1}^{(t)}(s', \pi_{h+1}^{(t_{\text{last}})}(s')) \right] \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[f_{h+1}^{(t)}(s', \pi_{h+1}^{(t)}(s')) - f_{h+1}^{(t)}(s', \pi_{h+1}^{(t_{\text{last}})}(s')) \right] \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\left\langle f_{h+1}^{(t)}(s', \cdot), \pi_{h+1}^{(t)}(\cdot|s') - \pi_{h+1}^{(t_{\text{last}})}(\cdot|s') \right\rangle \right].
\end{aligned} \tag{102}$$

We can now continue to go through the mirror descent argument. Recall that

$$\pi_{h+1}^{(t+1)}(\cdot|s') = \frac{\pi_{h+1}^{(t)}(\cdot|s') \exp(\eta f_{h+1}^{(t)}(s', \cdot))}{\sum_{a \in \mathcal{A}} \pi_{h+1}^{(t)}(a|s') \exp(\eta f_{h+1}^{(t)}(s', a))} = Z^{-1} \pi_{h+1}^{(t)}(\cdot|s') \exp(\eta f_{h+1}^{(t)}(s', \cdot)),$$

and rearranging this yields

$$\eta f_{h+1}^{(t)}(s', \cdot) = \log Z_t + \log \pi_{h+1}^{(t+1)}(\cdot|s') - \log \pi_{h+1}^{(t)}(\cdot|s'),$$

where $\log Z_t$ is

$$\log Z_t = \log \left(\sum_{a \in \mathcal{A}} \pi_{h+1}^{(t)}(a|s') \exp(\eta f_{h+1}^{(t)}(s', a)) \right) = \log \pi_{h+1}^{(t)}(\cdot|s') - \log \pi_{h+1}^{(t+1)}(\cdot|s') + \eta f_{h+1}^{(t)}(s', \cdot).$$

Noting that $\sum_{a \in \mathcal{A}} \left(\pi_{h+1}^{(t)}(\cdot|s') - \pi_{h+1}^{(t_{\text{last}}+1)}(\cdot|s') \right) = 0$, we can now bound that

$$\left\langle \eta f_{h+1}^{(t)}(s', \cdot), \pi_{h+1}^{(t)}(\cdot|s') - \pi_{h+1}^{(t_{\text{last}}+1)}(\cdot|s') \right\rangle$$

$$= \left\langle \log Z_t + \log \pi_{h+1}^{(t+1)}(\cdot|s') - \log \pi_{h+1}^{(t)}(\cdot|s'), \pi_{h+1}^{(t)}(\cdot|s') - \pi_{h+1}^{(t_{\text{last}}+1)}(\cdot|s') \right\rangle$$
$$= \left\langle \log \pi_{h+1}^{(t+1)}(\cdot|s') - \log \pi_{h+1}^{(t)}(\cdot|s'), \pi_{h+1}^{(t)}(\cdot|s') - \pi_{h+1}^{(t_{\text{last}}+1)}(\cdot|s') \right\rangle$$
$$= -\text{KL}\left(\pi_{h+1}^{(t)}(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right) + \text{KL}\left(\pi_{h+1}^{(t_{\text{last}}+1)}(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right) - \text{KL}\left(\pi_{h+1}^{(t_{\text{last}}+1)}(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right), \quad (103)$$

where the last equality follows directly from Lemma 24. This establishes the following telescoping sum

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\left\langle f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^{(t)}(\cdot|s') - \pi_{h+1}^{(t_{\text{last}}+1)}(\cdot|s')\right\rangle\right]$$
$$= \frac{1}{\eta}\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[-\text{KL}\left(\pi_{h+1}^{(t)}(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right) + \text{KL}\left(\pi_{h+1}^{(t_{\text{last}}+1)}(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right) - \text{KL}\left(\pi_{h+1}^{(t_{\text{last}}+1)}(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right)\right]$$
$$= -\frac{1}{\eta}\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\text{KL}\left(\pi_{h+1}^{(t)}(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right)\right]$$
$$+ \frac{1}{\eta}\sum_{h=1}^{H}\sum_{k=1}^{K_h}\sum_{t=t_k}^{t_{k+1}-1} \mathbb{E}_{\pi^*}\left[\text{KL}\left(\pi_{h+1}^{(t_k+1)}(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right) - \text{KL}\left(\pi_{h+1}^{(t_k+1)}(\cdot|s') \,||\, \pi_{h+1}^{(t)}(\cdot|s')\right)\right]$$
$$= -\frac{1}{\eta}\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\text{KL}\left(\pi_{h+1}^{(t)}(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right)\right] + \frac{1}{\eta}\sum_{h=1}^{H}\sum_{k=1}^{K_h} \mathbb{E}_{\pi^*}\left[\text{KL}\left(\pi_{h+1}^{(t_k+1)}(\cdot|s') \,||\, \pi_{h+1}^{(t_{k+1})}(\cdot|s')\right)\right]. \quad (104)$$

To see this step, consider the example where we perform switches at step 1, 3, 6, and $T = 8$. Note that we adopt the convention that $\pi^{(T+1)} = \pi^{(T)}$. The telescoping sum then becomes

$$\sum_{t=1}^{8} \text{KL}(t_{\text{last}}+1||t+1) - \text{KL}(t_{\text{last}}+1||t)$$
$$= \text{KL}(2,2) - \text{KL}(2,1) + \text{KL}(2,3) - \text{KL}(2,2) + \text{KL}(4,4) - \text{KL}(4,3) + \text{KL}(4,5) - \text{KL}(4,4)$$
$$+ \text{KL}(4,6) - \text{KL}(4,5) + \text{KL}(7,7) - \text{KL}(7,6) + \text{KL}(7,8) - \text{KL}(7,7) + \text{KL}(7,9) - \text{KL}(7,8)$$
$$= \text{KL}(2,3) + \text{KL}(4,6) + \text{KL}(7,9). \quad (105)$$

So the sum includes every t_k where a switch occurs:

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\left\langle f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^{(t)}(\cdot|s') - \pi_{h+1}^{(t_{\text{last}})}(\cdot|s')\right\rangle\right]$$
$$= \frac{1}{\eta}\sum_{h=1}^{H}\sum_{k=1}^{K_h} \mathbb{E}_{\pi^*}\left[\text{KL}\left(\pi_{h+1}^{(t_k+1)}(\cdot|s') \,||\, \pi_{h+1}^{(t_{k+1})}(\cdot|s')\right)\right] - \frac{1}{\eta}\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\text{KL}\left(\pi_{h+1}^{(t)}(\cdot|s') \,||\, \pi_{h+1}^{(t+1)}(\cdot|s')\right)\right]$$
$$= \mathcal{H}^*(\pi_h^{(t)}, t_k).$$

The latter two terms cancel in the TV distance, or any distance where the triangle inequality holds. It is harder to see a relation with the KL divergence, but in general, one may not be able to achieve sublinear regret.

This is because, if we merge this term with the first term in the regret decomposition of Lemma 8, we obtain

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\left\langle f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^{*}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')\right\rangle\right] + \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\left\langle f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^{(t)}(\cdot|s') - \pi_{h+1}^{(t_{\text{last}})}(\cdot|s')\right\rangle\right],$$

which evaluates to

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\left\langle f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^{*}(\cdot|s') - \pi_{h+1}^{(t_{\text{last}})}(\cdot|s')\right\rangle\right].$$

G. Miscellaneous Lemmas

In this section, we collect some auxiliary lemmas that are useful in deriving our main results.

Lemma 20 (Bound on Covering Number of Value Function Class, Lemma B.1 from Zhong and Zhang (2023)). *Consider the value function class induced by a Q-function class $\mathcal{F}^{(t)}$ and a class of stochastic policies $\Pi^{(t)}$, given by*

$$\mathcal{V}_h^{(t)} = \left\{ \langle f_h(s,\cdot), \pi_h(\cdot|s) \rangle \mid f_h \in \mathcal{F}_h^{(t)}, \pi \in \Pi^{(t)} \right\}.$$

Then, the covering number of the value function class can be bounded by the product of the covering number of its components:

$$\mathcal{N}_{\mathcal{V}_h^{(t)}}(\rho) \leq \mathcal{N}_{\mathcal{F}_h^{(t)}}(\rho/2) \cdot \mathcal{N}_{\Pi_h^{(t)}}(\rho/2H).$$

Lemma 21 (Lemma B.3, Zhong and Zhang (2023)). *For $\pi, \pi' \in \Delta(\mathcal{A})$ and $Q, Q' : \mathcal{A} \mapsto \mathbb{R}^+$, if $\pi(\cdot) \propto \exp(Q(\cdot))$ and $\pi'(\cdot) \propto \exp(Q'(\cdot))$, we have*

$$\|\pi - \pi'\|_1 \leq \sqrt{2 \cdot KL(\pi||\pi')} \leq 2\sqrt{\|Q - Q'\|_\infty}.$$

Lemma 22 (Adapted Version of Lemma D.2 of Xiong et al. (2023)). *Let $\mathcal{F}$ be a function class with low D_Δ-type Bellman Eluder dimension. Then, for any policy $\pi \in \Pi$, if we have that $\sum_{i=1}^{t-1}(f_h^{(t)} - \mathcal{T}_h^\pi f_{h+1}^{(t)})^2(s_h^{(i)}, a_h^{(i)}) \leq \beta H^2$ for any $t \in [T]$ and $\beta \geq 9$, then for any $t' \in [T]$ we also have that*

$$\sum_{i=1}^{t}(f_h^{(i)} - \mathcal{T}_h^\pi f_{h+1}^{(i)})^2 \left(s_h^{(i)}, a_h^{(i)}\right) \leq O\left(d_{BE}(\mathcal{F}, D_\Delta, 1/\sqrt{T})\beta H^2 \log T\right).$$

Lemma 23 (Value Difference/Generalized Policy Difference Lemma, (Cai et al., 2024; Efroni et al., 2020)). *Let π, π' be two policies and $f \in \mathcal{F}$ be any Q-function. Then for any $t \in [T]$ we have*

$$\begin{aligned} &f_1(s_1, \pi_1(s_1)) - V_1^{\pi'}(s_1) \\ &= \sum_{h=1}^{H} \mathbb{E}_{\pi'}\left[\langle f_h(s_h,\cdot), \pi_h(\cdot \mid s_h) - \pi'_h(\cdot \mid s_h)\rangle\right] + \sum_{h=1}^{H} \mathbb{E}_{\pi'}\left[f_h(s_h,a_h) - \mathcal{T}_h^{\pi'} f_{h+1}(s_h,a_h)\right]. \end{aligned}$$

Lemma 24. *For any probability distributions $\pi(\cdot), \pi_1(\cdot)$ and $\pi_2(\cdot)$ over space $\mathcal{S}$. We have following relationship holds:*

$$\left\langle \pi_1(\cdot) - \pi(\cdot), \log \pi(\cdot) - \log \pi_2(\cdot) \right\rangle = -\mathrm{KL}(\pi_1||\pi) + \mathrm{KL}(\pi_1||\pi_2) - \mathrm{KL}(\pi||\pi_2).$$

Proof. Note that for the first equality, we have

$$\begin{aligned} \left\langle \pi_1(\cdot) - \pi(\cdot), \log \pi(\cdot) - \log \pi_2(\cdot) \right\rangle &= \left\langle \pi_1(\cdot), \log \pi(\cdot) - \log \pi_2(\cdot) \right\rangle - \left\langle \pi(\cdot), \log \pi(\cdot) - \log \pi_2(\cdot) \right\rangle \\ &= \left\langle \pi_1(\cdot), \log \pi(\cdot) - \log \pi_1(\cdot) \right\rangle + \left\langle \pi_1(\cdot), \log \pi_1(\cdot) - \log \pi_2(\cdot) \right\rangle \\ &\quad - \left\langle \pi(\cdot), \log \pi(\cdot) - \log \pi_2(\cdot) \right\rangle \\ &= -\mathrm{KL}(\pi_1||\pi) + \mathrm{KL}(\pi_1||\pi_2) - \mathrm{KL}(\pi||\pi_2). \end{aligned}$$

□

Lemma 25 (Policy Optimization Difference). *Let $\pi^{(t)}, t = 1, ..., T$ be a sequence of policies updated by:*

$$\pi_{h+1}^{(t+1)}(\cdot \mid s') = \frac{\pi_{h+1}^{(t)}(\cdot \mid s') \exp\left(\eta f_{h+1}^{(t)}(s',\cdot)\right)}{\sum_{a \in \mathcal{A}} \pi_{h+1}^{(t)}(a \mid s') \exp\left(\eta f_{h+1}^{(t)}(s',a)\right)} = Z_t^{-1} \pi_{h+1}^{(t)}(\cdot \mid s') \exp\left(\eta f_{h+1}^{(t)}(s',\cdot)\right),$$

where $f^{(t)} \in \mathcal{F}$. For any $t_1, t_2 \in [T]$, where $\mu^{(t)}$ is an arbitrary set of distributions,

$$\sum_{t=\min\{t_1,t_2\}}^{\max\{t_1,t_2\}} \mathbb{E}_{\mu^{(t)}}\left[\left\langle f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^{(t_2)}(\cdot|s') - \pi_{h+1}^{(t_1)}(\cdot|s')\right\rangle\right] \le \eta H^2(|t_2 - t_1| + 1).$$

Proof. We observe that

$$\frac{\pi_{h+1}^{(t+1)}(\cdot \mid s')}{\pi_{h+1}^{(t)}(\cdot \mid s')} \ge \frac{\exp(0)}{|\mathcal{A}|^{-1}\sum_{a\in\mathcal{A}}\exp(\eta H)} \ge \exp(-\eta H),$$
$$\frac{\pi_{h+1}^{(t+1)}(\cdot \mid s')}{\pi_{h+1}^{(t)}(\cdot \mid s')} = \frac{\exp(\eta f_{h+1}^{(t)}(s',\cdot))}{\sum_{a\in\mathcal{A}}\pi_{h+1}^{(t)}(a \mid s')\exp(\eta f_{h+1}^{(t)}(s',a))} \le \frac{\exp(\eta H)}{|\mathcal{A}|^{-1}\sum_{a\in\mathcal{A}}\exp(0)} = \exp(\eta H).$$

So we can establish that since $e^{-x} \ge 1 - x$ for all $x \in \mathbb{R}$,

$$\begin{aligned}|\pi_{h+1}^{(t+1)}(\cdot|s') - \pi_{h+1}^{(t)}(\cdot|s')| &= \pi_{h+1}^{(t+1)}(\cdot|s') \cdot \left|1 - \frac{\pi_{h+1}^{(t)}(\cdot|s')}{\pi_{h+1}^{(t+1)}(\cdot|s')}\right| \le \pi_{h+1}^{(t+1)}(\cdot|s')\left(1 - \exp(-\eta H)\right) \\ &\le \pi_{h+1}^{(t+1)}(\cdot|s')\left(1 - (1 - \eta H)\right) \le \eta H \pi_{h+1}^{(t+1)}(\cdot|s'). \end{aligned} \tag{106}$$

By the triangle inequality and the fact that $f \le H$ for all $f \in \mathcal{F}$, it then follows that

$$\begin{aligned}&\sum_{t=\min\{t_1,t_2\}}^{\max\{t_1,t_2\}} \mathbb{E}_{\mu^{(t)}}\left[\left\langle f_{h+1}^{(t)}(s',\cdot), \pi_{h+1}^{(t_2)}(\cdot|s') - \pi_{h+1}^{(t_1)}(\cdot|s')\right\rangle\right] \le \sum_{t=\min\{t_1,t_2\}}^{\max\{t_1,t_2\}} \mathbb{E}_{\mu^{(t)}}\left[\left\langle H, \eta H \pi_{h+1}^{(t+1)}(\cdot|s')\right\rangle\right] \\ &\le \sum_{t=\min\{t_1,t_2\}}^{\max\{t_1,t_2\}} \eta H^2 \le 2\eta H^2(|t_2 - t_1| + 1). \end{aligned} \tag{107}$$

□

H. Further Experiment Details

Figure 2 can be reproduced by running `actor_critic.ipynb` within the following GitHub repository (https://github.com/hetankevin/hybridcov). Figure 3 can be reproduced by running `scripts/run_antmaze.sh` within the following GitHub repository (https://github.com/nakamotoo/Cal-QL). The results for Cal-QL arise from running the script as-is. Algorithm 2H can be reproduced by adding the flags `--enable_calql=False`, `--use_cql=False`, and `--online_use_cql=False`. Algorithm 1H can be reproduced with the same flags as Algorithm 2H, but additionally setting the `config.cql_max_target_backup` argument within the `ConservativeSAC()` object to False.

To implement the mirror descent update in Figure 2, we store the sequence of past Q-functions fitted for Algorithm 1, and the last Q-function for Algorithm 2. Upon receiving a query to evaluate or sample from $\pi_h^{(t)}(s,a)$ for a given h, s, a tuple, we compute $\exp(\eta \sum_{t=1}^{T} f_h^{(t)}(s,a))$ for Algorithm 1, and $\exp(\eta(t - t_{\text{last}}) f_h^{(t)}(s,a))$ for Algorithm 2. To generate a sample from this density, the normalizing constant can be computed exactly in the case where there is a finite number of actions. Otherwise a sample can be generated via MCMC, importance sampling, or rejection sampling.