A Theoretical Framework for Partially-Observed Reward States in RLHF

Chinmaya Kausik
University of Michigan

Mirco Mutti
Technion

Aldo Pacchiano
Broad Insitute of MIT & Harvard

Ambuj Tewari
University of Michigan

Abstract

The growing deployment of reinforcement learning from human feedback (RLHF) calls for a deeper theoretical investigation of its underlying models. The prevalent models of RLHF do not account for neuroscience-backed, partially-observed "internal states" that can affect human feedback, nor do they accommodate intermediate feedback during an interaction. Both of these can be instrumental in speeding up learning and improving alignment. To address these limitations, we model RLHF as reinforcement learning with partially observed reward-states (PORRL). We accommodate two kinds of feedback – cardinal and dueling feedback. We first demonstrate that PORRL subsumes a wide class of RL problems, including traditional RL, RLHF, and reward machines. For cardinal feedback, we present two model-based methods (POR-UCRL, POR-UCBVI). We give both cardinal regret and sample complexity guarantees for the methods, showing that they improve over naive history-summarization. We then discuss the benefits of a model-free method like GOLF with naive history-summarization in settings with recursive internal states and dense intermediate feedback. For this purpose, we define a new history aware version of the Bellman-eluder dimension and give a new guarantee for GOLF in our setting, which can be exponentially sharper in illustrative examples. For dueling feedback, we show that a naive reduction to cardinal feedback fails to achieve sublinear dueling regret. We then present the first explicit reduction that converts guarantees for cardinal regret to dueling regret. In both feedback settings, we show that our models and guarantees generalize and extend existing ones.

1 Introduction

As automated systems become more ubiquitous, the need to understand how to align their objectives with the needs of humans that interact with them has become increasingly important [15, 22, 35, 28]. The development and study of reinforcement learning from human feedback (RLHF) has been an important way of formalizing these problems and design methods for alignment [16, 56]. RLHF is concerned with the study of how to find a policy that maximizes an objective defined in terms of human labeled data in an RL domain [16, 56].

Many RLHF methods entail learning a reward function from human data, and then using the learned reward function as an input to a traditional reinforcement learning algorithm such as PPO [46]. These methods have been pivotal in the development of several technologies such as robotics [16, 7, 47], recommender systems [58], and most notably the training of large language models (LLMs) [6, 41, 1].

There exist two dominant kinds of feedback in reward-based RLHF, *cardinal* and *dueling* feedback. Cardinal feedback requires the human labeler to provide a single label over an entire trajectory of interaction between the agent and the environment [19, 11]. Dueling feedback requires the human to specify a preference between two trajectories. In practice, dueling feedback has been

Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2024).

used to train reward models, which have been successfully combined with RL algorithms to train LLMs [62, 48, 3, 6, 41]. Past work [11, 55, 45] has designed algorithms for both cardinal and dueling feedback under various metrics – standard/cardinal regret, sample complexity or dueling regret.

We observe that current models of reward-based RLHF assume a very specific model of non-Markovian rewards. Modeling rewards as non-Markovian is natural, since human responses to stimuli are known to be affected by partially-observed and evolving "internal states" [20]. For example, when a human reads a piece of text (possibly generated by an LLM), their assessment may oscillate between opposing sentiments in different parts of the text. Unfortunately, current models do not explicitly incorporate such "internal states" that affect rewards, and are limited to a specific linear model of rewards. While one can incorporate internal states using naive history-summarization, i.e. by treating the entire trajectory $\tau[h]$ so far as the state, we show below that better general algorithms can be designed with improved guarantees.

Additionally, current models assume that feedback is received only once at the end of an episode or pair of episodes. In many applications such as motion [34] and mathematical reasoning [52], correctly incorporating intermediate or "snippet-level" feedback can speed up learning as well as improve alignment. With this in mind, we ask the following questions:

How do we generalize the RLHF setting to incorporate internal states and intermediate feedback? What algorithms and guarantees can improve over naive history-summarization here?

Contributions:

- **Introducing PORRL:** In Section 2, we introduce PORRL, which generalizes current RLHF models to incorporate "internal states" and intermediate feedback.
- **Improving over naive history-summarization:** In Section 3.1, we design model-based optimistic algorithms, POR-UCRL and POR-UCBVI, achieving a regret of $\widetilde{\mathcal{O}}((poly(H,S,A) + p\sqrt{d_E d_C})\sqrt{T})$ and a sample complexity of $\widetilde{O}((poly(H,S,A)/\varepsilon^2 + p^2 d_E d_C/\varepsilon^2)$ under minimal assumptions.[1] The $poly(H,S,A)$ term would be exponential in H under naive history-summarization. We show that our guarantees subsume and improve over past results in RLHF.
- **Leveraging recursive structure on internal states:** In Section 3.2, we study the model-free algorithm GOLF, applied using history-summarization. We define a new "history-aware" notion of dimension, d_{HABE} and show that GOLF has regret $\widetilde{\mathcal{O}}(pH\sqrt{d_{\mathrm{HABE}} d_C T})$. We show using an example that when internal states have a recursive structure, our guarantee can be exponentially smaller than existing guarantees and guarantees for our model-based methods.
- **Reduction from Dueling to Cardinal PORRL:** We show that a naive blackbox reduction from dueling to cardinal PORRL always fails. We design a whitebox reduction from dueling PORRL to a large class of optimistic algorithms for cardinal PORRL. To the best of our knowledge, this is the first explicit reduction that converts guarantees for cardinal regret to dueling regret in RL.

1.1 Related Work

RLHF. RL with human preferences has a long history [2, 8, 44]. It has been successfully used in disparate domains such as robotics, games, and LLMs. The problem of learning from cardinal feedback has been theoretically studied in [19, 11]. Theoretical guarantees for utility-based preferential (dueling) feedback can be found in [40, 45, 14, 60]. The non-Markovian nature of the optimal policy under these RLHF models contributes greatly to why the problem is harder than traditional RL.

Internal states and intermediate feedback. There is evidence in neuroscience research indicating that human responses to stimuli are affected by "internal states" — partially hidden variables that profoundly shape perception, cognition, and action" [see 20]. Despite not explicitly recognizing the phenomenon of human internal states, several works in RLHF incorporate richer forms of feedback. For example, Wu et al. [57] consider human labeling over sub-sections of the text. In work on process supervision [52, 36], humans give feedback on intermediate steps. Motivated by these, our work is a first attempt at laying the groundwork for a theoretical treatment of internal human states and intermediate feedback in RLHF, using partially observed reward-states.

Partial observability in RL. The problem of partial observability in RL is not new. Although learning in POMDPs [4] is known to be statistically intractable in general [33, 31], a flurry of recent works

[1] d_E is a relevant eluder dimension and d_C is a relevant covering dimension.

have studied POMDPs under various structural assumptions [18, 37, 38, 21, 59, 9, 12, 13, 54, 61]. Our model is distinct from POMDPs since our results do not require the latent state evolution to be Markovian, but assumes Markovian transitions for observed states. See Section 2.3 for a discussion.

2 Defining RL with Partially-Observed Reward States (PORRL)

In this paper, we consider an episodic reinforcement learning setting in which a learner interacts with an MDP having a state space $\mathcal{S}$, an action space $\mathcal{A}$, transitions dynamics $\mathbb{P}$, and episode length H. At each time-step $h \in [H]$ of an episode, the learner observes the state s_h and takes an action a_h, generating a *trajectory* $\tau = (s_1, a_1, \cdots, s_H, a_H) \in \Gamma$, where Γ denotes the space of trajectories.[2] In a typical RLHF setting, the learner observes a human feedback $o_H \in \mathcal{O}$ at the end of the episode, which is associated to but potentially different from a reward $r : \Gamma \to \mathbb{R}$ encoding the task. We now describe how internal states and intermediate feedback shall be incorporated in the latter RLHF framework through a guiding example, and we use this to formally introduce the PORMDP model.

2.1 PORMDPs

Let us consider the example of a human interacting with a language model. In this example, an action is a token, the state is the text so far, and reward is some score representing the human's satisfaction, which induces stochastic feedback. The internal states could be the human's emotional reaction to the text (e.g., happy, frustrated, or amused), or numbers in $[0, 1]$ encoding a confidence level that the text is progressing towards a coherent response. While an agent goes through a sequence of states and actions, the system (i.e., the human) progresses through internal states, which inevitably affect, together with agent's actions and the state of the process, the human's satisfaction.

Formally, this can be modeled by introducing internal states $u \in \mathcal{U}$ and defining the set of *underlying* histories Γ^u_{h-1} that incorporate internal states by $\Gamma^u_{h-1} := \{\tau^u[h-1] = \{(s_l, u_l, a_l)\}_{l=1}^{h-1} \mid s_l \in \mathcal{S}, a_l \in \mathcal{A}, u_l \in \mathcal{U}\}$. We model the dynamics of internal states by saying that there exists an internal state generator $w_h : \Gamma^u_{h-1} \times \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{U})$ so that the human's internal state u_h is sampled from the distribution defined by $w_h(\tau^u[h-1], s_h, a_h)$. The human's satisfaction at time h should then be a function of the current state and action, but also the current internal state, given by $r_h(s_h, u_h, a_h)$.

The agent does not observe the reward r_h directly, but a feedback o_h depending on r_h. Typically, o_h will be $\{0, 1\}$ feedback reflecting whether the human says that they are satisfied or not. In general, this could be stochastic. For instance, this could be $Ber(\sigma_h(r_h))$ for some function σ_h. So, $o_h \sim e_h(r_h)$ for some distribution $e_h(r_h)$. This leads to the general definition below, where we have introduced new objects $\mathcal{U}, \mathcal{H}_p, w, e$ not seen in traditional RL:

Definition 1. A PORMDP $\mathbb{M}$ with *cardinal feedback* is a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{U}, \mathcal{O}, \mathbb{P}, \mathcal{H}_p, r, w, e)$, where:

- $\mathcal{S}, \mathcal{A}$ are fully observable states and actions, $\mathcal{U}$ are *unobserved internal reward-states*, $\mathcal{O}$ is a space of feedback, $\mathbb{P}(\cdot \mid s, a)$ is a Markovian transition matrix, $s_1 \in \mathcal{S}$ is an initial state.[3]
- $\mathcal{H}_p \subset [H]$ is a set of timesteps where reward and feedback is obtained with size $|\mathcal{H}_p| = p$.
- $r := \{r_h\}_{h \in \mathcal{H}_p}$ so that $r_h : \mathcal{S} \times \mathcal{U} \times \mathcal{A} \to \mathbb{R}$ are reward functions at time h.
- $w := \{w_h\}_{h \in \mathcal{H}_p}$ so that $w_h : \Gamma^u_{h-1} \times \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{U})$ are *internal state generators* that map underlying histories of (s, a, u) tuples to distributions over $\mathcal{U}$.
- $e := \{e_h\}_{h \in \mathcal{H}_p}$ are *feedback functions* so that the feedback $o_h \sim e_h(r_h)$ is sampled from an η_h-subgaussian distribution e_h with mean $\sigma_h(r_h)$ for some activation function $\sigma_h : \mathbb{R} \to \mathbb{R}$.[4]

In some relevant RLHF applications, the human is presented with two trajectories and they provide feedback based on the pair. In most cases, this involves indicating a 0-1 preference between trajectories. To accommodate this setting, we extend the framework to dueling feedback.

Definition 2. A PORMDP $\mathbb{M}$ with *dueling feedback* is a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{U}, \mathcal{O}, \mathbb{P}, \mathcal{H}_p, r, w, e)$, where everything is identical to Definition 1, except that every episode now involves running two trajectories τ_1, τ_2 that produce rewards $r_{h,1}, r_{h,2}$ $\forall h \in \mathcal{H}_p$, and feedback is distributed as $o_h \sim e_h(r_{h,1} - r_{h,2})$.

[2] We will further denote $\tau[h] = (s_1, a_1, \ldots, s_h, a_h)$ the sub-trajectory of τ of length h and Γ_h the corresponding space of sub-trajectories of length h.

[3] Recall that choosing a formal state s_1 to serve as a placeholder initial state is not restrictive.

[4] This subsumes and generalizes the example of Bernoulli feedback in RLHF.

We note that PORMDPs subsume and model a wide class of RL settings, including RLHF. A brief list of settings that PORMDPs subsume is as follows: (i) traditional MDPs, by setting $\mathcal{U} = \{\star\}$; (ii) existing linear models of RLHF, setting $\mathcal{U} = \{\phi(\tau)^\top \mathbf{w}\}$ for a known feature map ϕ and unknown $\mathbf{w}$ [11, 19, 45, 55]; (iii) learning reward models with stochastic feedback by setting $\mathcal{U}$ to be the set of reward states [26, 24, 25, 23]. By using $\mathcal{U}$ to model implicit intentions, PORMDPs can also model learning from the following feedback: (iv) process supervision [36, 52], (v) fine-grained feedback [57] and (vi) snippet-level feedback [34]. Further, one can show that in all these settings, we can define the $\mathcal{U}$ generators w_h to be deterministic.

One illustrative hard example of PORRL is that of a *combination lock,*[5] which we will also use later in the paper. Consider an H-digit numerical lock with a set $\mathcal{A}$ of options at each digit. Let the true combination be $a_1^\star, \dots a_H^\star$. An agent tries to unlock it by listening for "clicks" while rotating the dial at each digit h. Naturally, we only hear clicks at digit h if the entire combination so far is correct. We thus model this as a PORMDP with non-Markovian rewards, $\mathcal{S} = \{\star\}, \mathcal{U} = \{\bigcup_h \mathcal{A}^h\}$ and the appropriate dynamics. Arguing that the click might sometimes be too faint, we consider stochastic rewards. Specifically, we model this as $r_h(s_h, u_h, a_h) = Ber(q\mathbb{1}_{a_1^\star, \dots a_h^\star}(u_h))$ for some uncertainty parameter q. Notice that the internal states have a recursive structure here, and they evolve in a Markovian way. This is a toy model for the problem of learning to take desirable sequences of actions using intermediate feedback. It can be viewed as a simplified version of many such tasks – navigating mazes, writing structured essays with guidance, writing a proof with feedback on correctness.

2.2 Reinforcement Learning in PORMDPs (PORRL) with Cardinal and Dueling Feedback

Due to the complex nature of observability in our problem, we will use this subsection to carefully set up a meaningful set of RL problems, in which an agent interacts with a PORMDP to optimize a policy. At each step h, the agent observes a history $\tau[h-1] \in \Gamma_{h-1}$ and takes an action $a_h \sim \pi(\tau[h-1], s_h)$. The agent does not observe the reward r_h, but receives an observation $o_h \sim e_h(r_h)$.

Defining the learning objective. Since rewards are partially observed and dependent on the entire history, there is a subtlety in defining value functions. We first choose and fix some subclass Π of history-dependent policies and we define the total expected reward of a policy $\pi \in \Pi$ as

$$V_w(\mathbb{M}, \pi) := \mathbb{E}_{\tau^u \sim \mathbb{P}^{w,\pi}} \left[\sum_{h \in \mathcal{H}_p} r_h(s_h, u_h, a_h) \right]$$

$V_w(\mathbb{M}, \pi)$ is taking an expectation over the dynamics of *underlying* trajectories $\tau^u = \{(s_h, u_h, a_h)\}_{h=1}^H \sim \mathbb{P}^{w,\pi}$. Since the states u are never revealed, these dynamics can never be learnt, making V_w hard to directly deal with. In this light, we introduce stochastic functions $g_h : \Gamma_h \to \Delta(\mathcal{U})$ that marginalize the internal state generator w_h over the sequence $u_1, \dots u_{h-1}$. That is, given an (s, a) history $\tau[h]$, we can define[6] $g_h(\tau[h]) \sim u_h \mid \tau[h]$. Now define

$$V_g(\mathbb{M}, \pi) := \mathbb{E}_{\tau \sim \mathbb{P}^\pi} \left[\sum_{h \in \mathcal{H}_p} \mathbb{E}_{u_h \sim g_h(\tau[h])} [r_h(s_h, u_h, a_h)] \right]$$

$V_g(\mathbb{M}, \pi)$ is a much more tractable object, where the outer expectation is taken over the dynamics of the *observed* trajectories τ. The following result establishes that as one would hope, $V_w = V_g$.

Lemma 1 (Replacing w with g)**.** *For any history-dependent policy π that selects an action $a_h \sim \pi(\tau[h-1], s_h)$, $V_w(\mathbb{M}, \pi) = V_g(\mathbb{M}, \pi)$ holds for any $\mathbb{M}$.*

For the purposes of value functions, $\mathbb{M}$ is fully specified by $(\mathcal{S}, \mathcal{A}, \mathcal{U}, \mathcal{O}, \mathbb{P}, \mathcal{H}_p, r, g, e)$. Henceforth we replace w with g and denote the value function $V_g(\mathbb{M}, \pi)$ by $V(\mathbb{M}, \pi)$. Define the optimal policy as $\pi_\star := \arg\max_{\pi \in \Pi} V(\mathbb{M}, \pi)$.

Cardinal PORRL. Consider an algorithm producing a sequence of policies $\pi_1, \dots, \pi_T \in \Pi$, where π_t is chosen only using trajectories $\{\tau_i\}_{i=1}^{t-1}$ generated by $\{\pi_i\}_{i=1}^{t-1}$. We measure the performance of such an algorithm by its *cardinal regret* under model $\mathbb{M}_\star$:

$$\text{Regret}(T) = \sum_{t=1}^{T} V(\mathbb{M}_\star, \pi_\star) - V(\mathbb{M}_\star, \pi_t)$$

[5] This is a variant of a common example used to generate lower bounds in POMDPs [33, 31]. In contrast, we will use it to illustrate the power of our upper bounds.

[6] More technically, define $g_h(\tau[h])$ to be the *regular conditional distribution* of the random variable $w_h((\tau[h-1], u_1, \dots u_{h-1}), s_h, a_h)$, conditioned on $\tau[h]$.

One can also ask for the sample complexity of learning a good policy. Given a randomized algorithm that completes N episodes of interaction and outputs π_N, the *sample complexity* $N(\varepsilon, \delta)$ of the algorithm is the minimum N so that $V(\mathbb{M}_\star, \pi_\star) - V(\mathbb{M}_\star, \pi_N) \leqslant \varepsilon$ with probability at least $1 - \delta$ over the randomness of the feedback and the algorithm. It makes sense to study cardinal regret and sample complexity in two RLHF settings:

- Using a learnt reward model: In most deployments of offline RLHF, an offline dataset of dueling feedback from humans is typically used to create a cardinal feedback oracle (a reward model), which is then used to train the policy using RL. In fact, [43] do exactly this under our model. The sample complexity of the algorithm is important in this setting.
- Improving a deployed model with batched feedback: One can learn from batches of interaction with humans and hope to improve the model/policy adaptively over multiple batches. This is compatible with deploying LLMs or recommender systems to users, collecting a batch of good/bad feedback, and then fine-tuning the model offline using this batch. This approach is also discussed in [49, 17]. Regret is a better metric than sample complexity here, since we want users to be satisfied (exploiting) while improving the model (exploring). Instead of good/bad feedback, we can also ask for dueling feedback against a fixed policy π_0 and treat it as cardinal feedback.[7]

Dueling PORRL. In dueling PORRL, we play a *duel* by running two policies $(\pi_1, \pi_2) \in \Pi \times \Pi$ in parallel to obtain trajectories (τ_1, τ_2) and receive feedback $\{o_h\}_{h \in \mathcal{H}_p}$. Again, note that the rewards of the policies are not observed. While the definitions of $V(\mathbb{M}, \pi)$ and $\pi_\star$ are the same as before, we define a new measure of regret accordingly. If we play T duels $(\pi_{1,1}, \pi_{2,1}), \ldots, (\pi_{1,T}, \pi_{2,T})$ according to an algorithm, we aim to minimize the *dueling regret* given by

$$\mathrm{Regret}_D(T) = \sum_{t=1}^{T} V(\mathbb{M}_\star, \pi_\star) - \frac{V(\mathbb{M}_\star, \pi_{1,t}) + V(\mathbb{M}_\star, \pi_{2,t})}{2}$$

It makes sense to consider this metric when improving a deployed model with batched dueling feedback. We can do the same batching as the batched feedback example above, but instead compare our model/policy π_t to a fixed base policy π_0 and ask for dueling feedback. The induced feedback can be treated as cardinal feedback. This is similar to the ideas in [55], who consider this setting and give cardinal regret/sample complexity guarantees. However, when deploying a model, we typically want humans to be satisfied with *both* the options they are given. Cardinal regret only accounts for one of the options being good. Dueling regret demands that both policies used are good policies.

Remark 1. PORRL subsumes the settings of [45, 11], which in turn subsume the feedback models of RLHF [55]. Crucially, [55, 11] measure performance using only sample complexity or cardinal regret, while [45] only study dueling regret. We have discussed above why both metrics are important.

2.3 A General Yet Tractable Case

The nature of the feedback in PORMDPs , which depends on a reward that is function of the entire history, signals that PORRL may be intractable in general. We now instantiate the model into a statistically tractable sub-class that still subsumes most existing work on RLHF and all the examples provided at the end of Section 2.1. Specifically, we assume that the internal reward-state functions g_h are deterministic and the feedback is emitted according to a Bernoulli distribution depending on the reward. We will work under this assumption in the remainder of the paper.

Assumption 1. We work in a known class $\mathcal{P}$ of transition kernels $\mathbb{P}$ and known classes $\mathcal{R}_h$ of reward functions $r_h : \mathcal{S} \times \mathcal{U} \times \mathcal{A} \to \mathbb{R}$ with $|r_h| \leqslant B$ for all h. Let g_h be deterministic (but unknown) and belonging to a known class of "decoder functions" $\mathcal{G}_h$. Let $\mathcal{O} = \{0, 1\}$ and let e_h only depend on the rewards. For dueling feedback, let $e_h(r_{h,1} - r_{h,2})$ be η_h-subgaussian with known mean $\sigma_h(r_{1,h} - r_{2,h})$. Also assume that σ_h and σ_h^{-1} are Lipschitz with Lipschitz constants $\kappa_{1,h}$ and $\kappa_{2,h}$ respectively. Call the resulting class of PORMDPs $\mathcal{M}$.

We also define a function class induced by $\mathcal{R}_h$ and $\mathcal{G}_h$.

Definition 3. Let us then consider the decoder-induced function classes $\mathcal{F}_h$ given by

$$\mathcal{F}_h := \left\{ f_h : \Gamma_h \to \mathbb{R} \;\middle|\; \exists g_h \in \mathcal{G}_h, r_h \in \mathcal{R}_h \quad \text{s.t.} \quad f_h(\tau[h]) = r_h(s_h, g_h(\tau[h-1]), a_h),\ \forall \tau \right\}$$

[7] If the activation function is Lipschitz and monotone, then we can get cardinal regret guarantees for this problem by using the difference function class.

Also define $\mathcal{F} := \prod_{h\in\mathcal{H}_p} \mathcal{F}_h$ so that $f = \{f_h\}_{h\in\mathcal{H}_p} \in \mathcal{F}$. A model $\mathbb{M}$ is then fully determined by $(\mathbb{P}, f)$, so we denote $V(\mathbb{P}, f, \pi) := V(\mathbb{M}, \pi)$. Note that $V(\mathbb{P}, f, \pi) = \mathbb{E}_{\tau\in\mathbb{P}^\pi}\left[\sum_{h\in\mathcal{H}_p} f_h(\tau[h])\right]$.

Remark 2. We note that all examples from Section 2.1 work with deterministic dynamics for $\mathcal{U}$ and satisfy Assumption 1.

Giving statistically efficient algorithms for this framework comes with numerous challenges:

- **Traditional RL incurs linear regret:** We show in Lemma 3 that any method returning a possibly time-dependent but memoryless policy can incur linear regret.
- **POMDP results do not apply:** PORMDPs cannot be viewed as a subcase of POMDPs with latent states $\mathcal{S} \times \mathcal{U}$ since $s, u, a \to s', u'$ is not Markovian.[8] Even if we considered the subclass of PORMDPs where $s, u, a \to s', u'$ is Markovian, which would be a subclass of reward machines, this is a specific kind of overcomplete POMDP . Literature on overcomplete POMDPs is much more scarce than their undercomplete counterpart. The only paper that gives guarantees for overcomplete POMDPs to our knowledge is [37], which makes reward function *fully* observable. This cannot apply to our setting, since our rewards have to be *partially* observable.
- **Naive history-summarization is inefficient:** It is overkill to use naive history-summarization – where one treats the history $\tau[h]$ as the state s_h and executes traditional RL. This is because while policies are non-Markovian, state transitions are Markovian. It is unclear if we can leverage this structure without running into exponential dependence on H. Moreover, a lot of work on RL assumes that the reward is known, while learning the reward function is itself a hard problem here.
- **Ensuring satisfactory utilization of additional structure:** Examples like the combination lock signal that there are intuitive ways to leverage a recursive structure on the internal states. In the combination lock, one should wait for the "click" at each digit before moving onto the next digit, giving us a polynomial dependence on A, H in sample complexity. It is unclear if *general* algorithms for PORRL can implicitly leverage such structure to achieve polynomial guarantees.

3 Optimistic Algorithms for Cardinal PORRL

3.1 Improving over Naive History-Summarization with Model-Based Methods

In this section, we present two optimistic methods that leverage Markovian transitions in PORMDPs – POR-UCRL and POR-UCBVI. We describe them briefly here, deferring details to Appendix D, E.

- **POR-UCRL:** At each timestep t, we maintain a least squares estimate $\hat{f}^{t+1}$ of f and an MLE estimate $\hat{\mathbb{P}}_t$ and define confidence sets $\mathcal{C}_h^t(\delta)$ that consider all f_h with a small mean squared error against $\hat{f}_h^{t+1}$, such that $\mathcal{C}_{\mathcal{F}}^t(\delta) = \prod_{h=1}^H \mathcal{C}_h^t(\delta)$. The probability transition confidence sets $\mathcal{C}_{\mathcal{P}}^t(\delta)$ are the exact same ones in UCRL [27]. At timestep t, following confidence-set optimism, we play an optimistic policy $\widetilde{\pi}_t$ that maximizes its highest value $V(\mathbb{M}, \pi)$ over all models $\mathbb{M} \in \mathcal{C}_{\mathcal{F}}^t \times \mathcal{C}_{\mathcal{P}}^t$.
- **POR-UCBVI:** It is trickier to adapt ideas from UCBVI [5]. Yet again, we maintain a least squares estimate $\hat{f}^t$ and an MLE estimate $\hat{\mathbb{P}}_t$. Instead of confidence sets, we design trajectory-dependent bonuses for $\mathcal{F}$ as $b_{\mathcal{F}}^t(\tau, \delta) = \sum_{h\in\mathcal{H}_p} \max_{f_h, f_h' \in \mathcal{C}_h^t(\delta)} f_h(\tau[h]) - f_h'(\tau[h])$. We use these to define policy-level bonuses for $\mathcal{F}$ as $b_{\mathcal{F}}^t(\mathbb{P}, \pi, \delta) := \mathbb{E}_{\tau\sim\mathbb{P}_\star^\pi}\left[b_{\mathcal{F}}^t(\tau, \delta)\right]$. Then, we use the standard UCBVI bonuses to similarly define policy level bonuses for $\mathcal{P}$. At timestep t, following bonus-based optimism, we play an optimistic policy $\tilde{\pi}_t$ that maximizes its bonus-boosted value under $\hat{f}^t, \hat{\mathbb{P}}^t$.

We show that POR-UCRL enjoys the guarantee below.

Theorem 1 (POR-UCRL Regret)**.** *Under Assumption 1, the regret* $\mathrm{Regret}(T)$ *of POR-UCRL is bounded by the following with probability at least* $1-\delta$

$$\widetilde{\mathcal{O}}\left(\left(pS\sqrt{HA} + \sum_{h\in\mathcal{H}_p} \sqrt{d_{E,h} d_{C,h}}\right)\sqrt{T}\right)$$

where $d_{E,h} = \dim_E\left(\mathcal{F}_h, \frac{B}{T}\right)$ *and* $d_{C,h} = \log(\mathcal{N}(\mathcal{F}_h, 1/T, \|\cdot\|_\infty))$.

Here, the first term comes from uncertainty in $\mathbb{P}$. Under naive history-summarization, the first term would be exponential in H since the modified state space of trajectories would have size $\Omega((SA)^H)$.

[8] Since observed state transitions are Markovian, PORMDPs are also not more general than POMDPs .

Similar regret guarantees are given for POR-UCBVI in Theorem 8. Both guarantees are proved by viewing each algorithm as a specific instance of a generic optimistic algorithm for PORRL (see Appendix C, D, E). By a simple regret-to-PAC conversion, we also show that POR-UCRL has sample complexity of $\widetilde{\mathcal{O}}\left(\frac{p^2HS^2A}{\varepsilon^2} + \frac{p^2 d_E d_C}{\varepsilon^2}\right)$, where $d_E := \max_{h\in\mathcal{H}_p} d_{E,h}$, and $d_C := \max_{h\in\mathcal{H}_p} d_{C,h}$. POR-UCBVI has sample complexity $\widetilde{\mathcal{O}}\left(\frac{p^2HSA\max(H,S)}{\varepsilon^2} + \frac{p^2 d_E \max(d_C,H)\log(1/\delta)}{\varepsilon^2}\right)$.

Challenges: There are three main technical challenges in proving these guarantees. First, we have to handle non-Markovian reward functions with Markovian transitions. Second, in POR-UCBVI, we have the added challenge of ensuring that the bonus is uniformly optimistic over all history-dependent policies. This is typically a doubly exponential set ($A^{(SA)^H}$), so a union bound does not help us. Third, we are working with general function approximation for reward functions.

Remark 3 (Comparison to Past Results)**.** Notice that with $\mathcal{U} = \phi(\tau)^\top \mathbf{w}$ with $\mathbf{w} \in \mathbb{R}^d$ and $\mathcal{H}_p = \{H\}$, we are in the setting of [11]. Here, $d_{E,H} = d_{C,H} = d$, so POR-UCRL improves over their guarantees. POR-UCBVI improves over their guarantees by a smaller amount. With respect to sample complexity guarantees, we compare to [55]. While they use dueling feedback, our methods use cardinal feedback. In their setting, $\mathcal{U}$ is the set of all histories and $\mathcal{H}_p = \{H\}$. Their best guarantee is from P-OMLE, which makes $\widetilde{\mathcal{O}}\left(\frac{H^2S^2A}{\varepsilon^2} + \frac{H^2 d_{E,H} d_{C,H}}{\varepsilon^2}\right)$ dueling oracle queries for tabular $\mathcal{P}$. Both POR-UCRL and POR-UCBVI have a smaller complexity for cardinal feedback queries.

3.2 Leveraging Recursive Structures Using Model-Free Methods

We have established that the model-based methods POR-UCRL and POR-UCBVI improve over naive history-summarization and have a $poly(S, A, H)$ guarantee in terms of transition function estimation. However, we recall the last challenge mentioned in Section 2.3 – can they adapt to examples like the combination lock, where there is a recursive structure on the internal states? Disappointingly, we will see in Proposition 1 that the answer is no, since the eluder dimension of reward functions is A^h for the combination lock. Since POR-UCRL and POR-UCBVI decouple the learning of reward functions across timesteps, they are unable to incorporate a recursive structure on the reward functions.

In this light, we consider model-free methods. Unlike model-based methods that have to account for Markovian transitions, we can simply use naive history-summarization here and treat $\tau[h]$ as the state for Q-functions Q_h. However, under history-summarization, there is a subtlety involved in choosing the class $\mathcal{Q}$ of Q-functions given a known class $\mathcal{M}$ of models. Using product classes $\mathcal{Q}_1 \times \cdots \times \mathcal{Q}_H$ is wasteful, since often exponentially many tuples in a product class cannot be realized by any model $\mathcal{M}$.[9] Instead, one should consider the class of all individual tuples $(Q_1, \ldots, Q_H)$ that can be realized by a model $\mathcal{M}$. In practice, this translates to the problem of good representation learning – one should use a shared network for all Q-functions instead of using a different network for each timestep. This is reflected in the experimental choices of [43].

Model-free methods rely on the Bellman error, which relates consecutive Q-functions and couples their learning. It is thus natural to expect model-free methods like GOLF [32] to adapt to a recursive structure on internal states and perform better than model-based methods. However, existing guarantees do not reflect this, since it turns out from Proposition 1 below that the Bellman eluder (BE) dimension of the combination lock problem with the minimal Q-function class is still A^H.

The issue is that the proof of GOLF bounds the h-step Bellman errors in a decoupled manner, which is why it still fails to incorporate a recursive structure on internal states. Intuitively, one wants to *wait* for Bellman errors at timesteps $1, \ldots, h-1$ to become small before bounding the Bellman error at h. In this light, given a parameter α, we define the function class

$$\mathcal{Q}(\alpha, h) := \left\{Q \in \mathcal{Q} \mid |\mathbb{E}_{\mu_l(Q)}[Q_l - \mathcal{T}_l Q_{l+1}]| \leqslant \alpha,\ \forall 1 \leqslant l \leqslant h\right\}$$

that considers all tuples $(Q_1, \ldots, Q_H)$ where the Bellman errors until step h are already low. We can use this class to define the α-history aware Bellman eluder dimension (HABE) of $\mathcal{Q}$ as follows. Recall that π_Q is the policy that acts greedily according to $Q = (Q_1, \ldots Q_H)$.

Definition 4. Consider the Bellman errors $\Phi_h := \left\{Q_h - \mathcal{T}_h Q_{h+1} \middle| Q \in \mathcal{Q}(\alpha, h-1)\right\}$. Denote $\mu_h(Q)$ the distribution induced on $\tau[h-1], a_h$ by π_Q and let $\mathcal{D}_{h,\mathcal{Q}} := \{\mu_h(Q) \mid Q \in \mathcal{Q}\}$. Let $\dim_{DE}$ the distributional eluder dimension and define $\dim_{\mathrm{HABE}}(\mathcal{Q}, \alpha, \varepsilon) := \max_h \dim_{DE}(\Phi_h, \mathcal{D}_{h,\mathcal{Q}(\alpha,h-1)}, \varepsilon)$.

[9] The reader can use the example of the combination lock to convince themselves of this.

Intuitively, α-HABE dimension measures how hard it is to reduce the Bellman error at timestep h if the errors at *previous timesteps* $1, \ldots, h-1$ are already small. This captures the hardness of adapting to the recursive structure on internal states one/a few timesteps at a time. We discuss in Appendix F.1 how the α-HABE dimension compares to the Bellman eluder dimension in general. We now give a new guarantee for GOLF using the α-HABE dimension.

Theorem 2 (Modified GOLF Regret)**.** *Let Assumption 1 hold, and let* $d_{\mathrm{HABE}} = \dim_{\mathrm{HABE}}(\mathcal{Q}, \alpha, \min(\alpha, \sqrt{1/T}))$*. Choose hyperparameter* $\beta = c\log(HT\mathcal{N}(\mathcal{Q} \cup \mathcal{G}, 1/T, \|\cdot\|_\infty))$ *for some universal constant* c *and the auxiliary function class* $\mathcal{G}$ *used in GOLF, and define* $d_{C,\mathcal{Q}} := \log(\mathcal{N}(\mathcal{Q} \cup \mathcal{G}, 1/T, \|\cdot\|_\infty))$*. Then, GOLF satisfies* $\mathrm{Regret}(T) = \mathcal{O}\left(pH\sqrt{d_{\mathrm{HABE}} d_{C,\mathcal{Q}} T}\right)$.

Using a regret-to-PAC conversion, we also show in Corollary 6 that the sample complexity of GOLF is $\widetilde{\mathcal{O}}\left(\frac{p^2 H^2 d_{\mathrm{HABE}} d_{C,\mathcal{Q}}}{\varepsilon^2}\right)$. As foreshadowed above, we now show in Proposition 1 that these guarantees can be polynomial even when the the usual guarantees for GOLF as well as guarantees for our model-based algorithms are exponential. Note that this improvement is achieved only given dense intermediate feedback. Under sparse intermediate feedback, one cannot adapt to internal states "a few timesteps at a time."

Proposition 1 (Dimensions for the Combination Lock)**.** *Consider the combination lock problem with model class* $\mathcal{M} = \mathcal{P} \times \mathcal{F}$ *and induced Q-function class* $\mathcal{Q}$*.*

- *Under dense intermediate feedback with* $\mathcal{H}_p = [H]$*, the* $\dim_{\mathrm{HABE}}(\mathcal{Q}, \alpha)A$ *for all* $\alpha < q$*, while its BE dimension is at least* $A^H - 2$*. The eluder dimension for reward functions* $\dim_E(\mathcal{F}_h, \frac{B}{T})$ *is at least* A^h *for any* $h \leqslant H$*.*
- *For sparse intermediate feedback with* $\mathcal{H}_p = \{H\}$ *and any* $\alpha > 0$*, the* α-HABE *dimension, the BE dimension and the eluder dimension of* $\mathcal{F}_H$ *are all at least* $A^H - 2$*.*

We discuss in Appendix F.1 that in general, we do not have an inequality in either direction. However, the α-HABE dimension is typically smaller.

4 Dueling to Optimism Reduction

The dueling and cardinal feedback models are intimately related. It is thus tempting to use algorithms for cardinal PORRL to solve dueling PORRL. However, we detail why the "obvious" reduction from dueling feedback to cardinal feedback fails. This both demonstrates the hardness of the problem and motivates our reduction.

4.1 The Naive Reduction Always Fails

Consider a modified PORMDP $\overline{\mathbb{M}}$ with $\overline{\mathcal{S}} := \mathcal{S} \times \mathcal{S}$, $\overline{\mathcal{A}} := \mathcal{A} \times \mathcal{A}$, $\overline{\mathbb{P}} := \mathbb{P} \otimes \mathbb{P}$, where we run the pair of policies $\overline{\pi} := (\pi_1, \pi_2)$ and obtain observations based on the decoder-induced function $\overline{f}_h(\tau_1[h], \tau_2[h]) := f_h(\tau_1[h]) - f_h(\tau_2[h])$. Consider the space of all such PORMDPs induced by $\mathcal{M}$, and denote it by $\overline{\mathcal{M}}$. Since cardinal feedback in $\overline{\mathbb{M}}$ exactly corresponds to dueling feedback in $\mathbb{M}$, it is tempting to restrict to searching over $\Pi \times \Pi$ and run any algorithm for cardinal PORRL on this modified PORMDP $\overline{\mathbb{M}}$ to achieve low dueling regret.

This fails because the feedback model and regret metric are fundamentally non-aligned in dueling feedback, unlike in cardinal feedback. While the agent receives dueling feedback over the duel for $(\pi_{1,t}, \pi_{2,t})$, dueling regret is instead concerned with duels for $(\pi_\star, \pi_{1,t})$ and $(\pi_\star, \pi_{2,t})$. Running an algorithm for cardinal PORRL on the modified MDP will maximize the dueling *feedback* itself. This is achieved by playing one good and one really bad policy, unlike the two good policies needed for low dueling regret. We formalize this in Lemma 4, showing that the naive reduction leads to linear dueling regret for *any* PORMDP and *any* cardinal PORRL algorithm with sublinear regret.

4.2 Reducing Dueling to Optimistic Cardinal PORRL

The naive reduction fails because maximizing dueling feedback can lead to bad policies being played. In this subsection, we present a white-box reduction where we ensure that we only play potentially good policies for *both* $\pi_{1,t}$ and $\pi_{2,t}$. We detail here how we can obtain an algorithm for the dueling feedback problem from *any* optimistic algorithm for cardinal PORRL. We will focus on the case of confidence sets here for smoother exposition, the much harder case of bonuses is treated in Appendix G.2. A *generic optimistic algorithm using confidence sets* maintains confidence sets $\mathcal{C}_{\mathcal{M}}(\mathcal{D}_t, \delta)$ using the collected dataset $\mathcal{D}_t$ of trajectories and feedback. We define it formally in

Appendix C.1. For the reduction to work, we require that the confidence sets are well-designed, as demanded by Assumption 2. This assumption is satisfied for confidence sets used by POR-UCRL.

Assumption 2 (Controlling Value Error due to Confidence Sets)**.** $\mathbb{M}_\star \in \mathcal{C}_{\mathcal{M}}(\mathcal{D}_t, \delta)$ for arbitrary sequences $(\mathbb{P}_t, f^t) \in \mathcal{C}_{\mathcal{M}}(\mathcal{D}_t, \delta)$, both $\left|\sum_{t=1}^T V(\mathbb{P}_t, f^t, \pi_t) - V(\mathbb{P}_\star, f^t, \pi_t)\right| = \widetilde{\mathcal{O}}(C_P(\mathcal{M}, T, \delta))$ and $\left|\sum_{t=1}^T V(\mathbb{P}_\star, f^t, \pi_t) - V(\mathbb{P}_\star, f_\star, \pi_t)\right| = \widetilde{\mathcal{O}}(C_F(\mathcal{M}, T, \delta))$ hold with probability $1 - \delta/2$ each.

The key insight is to use confidence sets from cardinal PORRL to search for $\pi_{1,t}$ and $\pi_{2,t}$ only among policies that *both* have a chance of being optimal. Then one plays the *most uncertain* duel among all possible choices for $\pi_{1,t}$ and $\pi_{2,t}$. This generalizes and abstracts out ideas in [42], which presents a specific algorithm to achieve low dueling regret in their model. We present the reduction to optimism over confidence sets in Algorithm 1, the version for bonuses is in Appendix G.2. Define $V_D(\overline{\mathbb{M}}, \pi, \pi') = V(\mathbb{M}, \pi) - V(\mathbb{M}, \pi')$. We compute the confidence sets $\mathcal{C}_{\overline{\mathcal{P}}}(\mathcal{D}, \delta)$ as the image of $\mathcal{C}_{\mathcal{P}}(\mathcal{D}, \delta)$ under $\mathbb{P} \mapsto \overline{\mathbb{P}}$. We compute $\mathcal{C}_{\overline{\mathcal{F}}}(\mathcal{D}, \delta)$ by treating $\{o_h\}_{h \in \mathcal{H}_p}$ as cardinal feedback in $\overline{\mathbb{M}}$. As an example, for POR-UCRL, we perform a least squares fit for $\overline{f}$ and use Lemma 7 to define our confidence sets again.

Algorithm 1 Reduction from Dueling to Cardinal Confidence-Set Optimism

Input Known reward function $\{r_h\}_{h=1}^H$, method to compute $\mathcal{C}_{\overline{\mathcal{M}}}(\mathcal{D}, \delta) \leftrightarrow \mathcal{C}_{\overline{\mathcal{P}}}(\mathcal{D}, \delta) \times \mathcal{C}_{\overline{\mathcal{F}}}(\mathcal{D}, \delta)$
Initialize dataset $\mathcal{D}_1 \leftarrow \{\}$, $\mathcal{C}_{\overline{\mathcal{M}}}(\mathcal{D}_1, \delta) := \overline{\mathcal{P}} \times \overline{\mathcal{F}}$
for $t = 1, ..., T$ **do**
  Compute $\Pi_t = \left\{\pi \in \Pi \middle| \exists \overline{\mathbb{M}} \in \mathcal{C}_{\overline{\mathcal{M}}}(\mathcal{D}_t, \delta) \text{ s.t. } V(\overline{\mathbb{M}}, \pi, \pi_1) \geqslant 0 \ \forall \pi_1 \in \Pi\right\}$ {Candidates $\pi_\star$}
  Play $(\pi_{1,t}, \pi_{2,t}) \in \arg\max_{\pi, \pi' \in \Pi_t} \max_{\overline{\mathbb{M}}, \overline{\mathbb{M}}' \in \mathcal{C}_{\overline{\mathcal{M}}}(\mathcal{D}_t, \delta)} V_D(\overline{\mathbb{M}}, \pi, \pi') - V_D(\overline{\mathbb{M}}', \pi, \pi')$ {Most uncertain duel}
  Observe trajectories $\tau_{i,t} = \left\{(s_{i,h}^t, a_{i,h}^t)\right\}_{h=1}^H$ along with feedback $\{o_h\}_{h \in \mathcal{H}_p}$
  Update $\mathcal{D}_t$ to $\mathcal{D}_{t+1}$ using the data and compute $\mathcal{C}_{\overline{\mathcal{P}}}(\mathcal{D}_{t+1}, \delta)$, $\mathcal{C}_{\overline{\mathcal{F}}}(\mathcal{D}_{t+1}, \delta)$
end for

We then get the following regret guarantee and the corresponding corollary.

Theorem 3 (Reduction from Dueling to Confidence-Set-Based Optimism)**.** *If the confidence sets $\mathcal{C}_{\mathcal{M}}(\mathcal{D}_t, \delta)$ satisfy Assumption 2, then the dueling regret $\mathrm{Regret}_D(T)$ of Algorithm 1 is given by*

$$\mathrm{Regret}_D(T) = \widetilde{\mathcal{O}}(C_P(\mathcal{M}, T, \delta) + C_F(\overline{\mathcal{M}}, T, \delta))$$

Note that complexity parameter C_F depends on $\overline{\mathcal{M}}$. It is a priori unclear how the complexity of $\overline{\mathcal{M}}$ relates to that of $\mathcal{M}$. Fortunately, Lemma 2 below settles this, and we can then use our results for POR-UCRL to get Corollary 1 below.

Lemma 2 (Relating $\mathcal{F}$ and $\overline{\mathcal{F}}$)**.** *For any function class $\mathcal{F}$, $\dim_E(\overline{\mathcal{F}}, \varepsilon) \leqslant 9 \dim_E(\mathcal{F}, \varepsilon/2)$.*

Corollary 1 (Dueling Regret using POR-UCRL Confidence Sets)**.** *The confidence sets from POR-UCRL satisfy Assumption 2 and using them in Algorithm 1 leads to the following regret bound $\mathrm{Regret}_D(T) = \widetilde{\mathcal{O}}\left(\left(pS\sqrt{HA} + \sum_{h \in \mathcal{H}_p} \sqrt{d_{E,h} d_{C,h}}\right)\sqrt{T}\right)$.*

5 Conclusions and Future Work

In this work, we have introduced PORMDPs and their analysis as a way to better model internal states of humans and intermediate feedback in RLHF. We have introduced two statistically efficient algorithms for handling partially observed reward-states and have shown that they improve over naive history summarization. We have noted that these methods subsume as well as improve over a lot of past work in RLHF. We have studied how one can further leverage a recursive structure over internal states using model-free methods. For this purpose, we have defined a new notion of dimension, the α-HABE dimension, that captures the hardness of utilizing the recursive structure. Finally, we have also provided a novel reduction from dueling regret to optimistic algorithms for cardinal regret. We hope that our ideas lay the groundwork for further theoretical understanding of the statistical limits of learning good policies when interacting with "stateful" feedback such as that of humans. Our algorithms and proofs are presented in high generality and modularity in the appendix, and we hope that they can be used to provide novel algorithms and bounds in the future.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Riad Akrour, Marc Schoenauer, and Michèle Sebag. April: Active preference learning-based reinforcement learning. In *European Conerence on Machine Learning and Knowledge Discovery in Databases*, 2012.

[3] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

[4] Karl Johan Åström. Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205, 1965.

[5] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, 2017.

[6] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[7] Daniel Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International Conference on Machine Learning*, 2019.

[8] Róbert Busa-Fekete and Eyke Hüllermeier. A survey of preference-based online learning with bandit algorithms. In *Algorithmic Learning Theory*, 2014.

[9] Qi Cai, Zhuoran Yang, and Zhaoran Wang. Reinforcement learning from partial observation: Linear function approximation with provable sample efficiency. In *International Conference on Machine Learning*, 2022.

[10] Jeffrey Chan, Aldo Pacchiano, Nilesh Tripuraneni, Yun S Song, Peter Bartlett, and Michael I Jordan. Parallelizing contextual linear bandits. *arXiv preprint arXiv:2105.10590*, 2021.

[11] Niladri Chatterji, Aldo Pacchiano, Peter Bartlett, and Michael Jordan. On the theory of reinforcement learning with once-per-episode feedback. In *Advances in Neural Information Processing Systems*, 2021.

[12] Fan Chen, Yu Bai, and Song Mei. Partially observable rl with b-stability: Unified structural condition and sharp sample-efficient algorithms. In *International Conference on Learning Representations*, 2022.

[13] Fan Chen, Huan Wang, Caiming Xiong, Song Mei, and Yu Bai. Lower bounds for learning in revealing pomdps. In *International Conference on Machine Learning*, 2023.

[14] Xiaoyu Chen, Han Zhong, Zhuoran Yang, Zhaoran Wang, and Liwei Wang. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*, 2022.

[15] Brian Christian. *The alignment problem: Machine learning and human values*. WW Norton & Company, 2020.

[16] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 2017.

[17] Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*, 2024.

[18] Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, 2019.

[19] Yonathan Efroni, Nadav Merlis, and Shie Mannor. Reinforcement learning with trajectory feedback. In *AAAI conference on artificial intelligence*, 2021.

[20] Steven W Flavell, Nadine Gogolla, Matthew Lovett-Barron, and Moriel Zelikowsky. The emergence and influence of internal states. *Neuron*, 2022.

[21] Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Learning in observable pomdps, without computationally intractable oracles. In *Advances in Neural Information Processing Systems*, 2022.

[22] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

[23] Rodrigo Andrés Toro Icarte. *Reward Machines*. PhD thesis, University of Toronto, Canada, 2022.

[24] Rodrigo Toro Icarte, Toryn Klassen, Richard Valenzano, and Sheila McIlraith. Using reward machines for high-level task specification and decomposition in reinforcement learning. In *International Conference on Machine Learning*, pages 2107–2116. PMLR, 2018.

[25] Rodrigo Toro Icarte, Toryn Q Klassen, Richard Valenzano, and Sheila A McIlraith. Reward machines: Exploiting reward function structure in reinforcement learning. *Journal of Artificial Intelligence Research*, 73:173–208, 2022.

[26] Rodrigo Toro Icarte, Ethan Waldie, Toryn Klassen, Rick Valenzano, Margarita Castro, and Sheila McIlraith. Learning reward machines for partially observable reinforcement learning. In *Advances in Neural Information Processing Systems*, 2019.

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

[28] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

[29] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low bellman rank are pac-learnable, 2016.

[30] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, 2018.

[31] Chi Jin, Sham Kakade, Akshay Krishnamurthy, and Qinghua Liu. Sample-efficient reinforcement learning of undercomplete pomdps. In *Advances in Neural Information Processing Systems*, 2020.

[32] Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. In *Advances in Neural Information Processing Systems*, 2021.

[33] Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Pac reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, 2016.

[34] Kimin Lee, Laura Smith, and Pieter Abbeel. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091*, 2021.

[35] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

[36] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

[37] Qinghua Liu, Alan Chung, Csaba Szepesvári, and Chi Jin. When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, 2022.

[38] Qinghua Liu, Praneeth Netrapalli, Csaba Szepesvari, and Chi Jin. Optimistic mle–a generic model-based algorithm for partially observable sequential decision making. *arXiv preprint arXiv:2209.14997*, 2022.

[39] Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, 2021.

[40] Ellen Novoseller, Yibing Wei, Yanan Sui, Yisong Yue, and Joel Burdick. Dueling posterior sampling for preference-based reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, 2020.

[41] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

[42] Aldo Pacchiano, Aadirupa Saha, and Jonathan Lee. Dueling rl: reinforcement learning with trajectory preferences. *arXiv preprint arXiv:2111.04850*, 2021.

[43] Anonymous Authors (Preprint). Enhancing multi-step reasoning via process-supervised reinforcement learning from human feedback. `https://openreview.net/pdf?id=3pIeOI1yPG`, 2024. [Accessed 21-05-2024].

[44] Dorsa Sadigh, Anca D Dragan, Shankar Sastry, and Sanjit A Seshia. Active preference-based learning of reward functions. Technical report, EECS Berkeley, 2017.

[45] Aadirupa Saha, Aldo Pacchiano, and Jonathan Lee. Dueling rl: Reinforcement learning with trajectory preferences. In *International Conference on Artificial Intelligence and Statistics*, 2023.

[46] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[47] Daniel Shin, Anca D Dragan, and Daniel S Brown. Benchmarks and algorithms for offline preference-based reward learning. *arXiv preprint arXiv:2301.01392*, 2023.

[48] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, 2020.

[49] Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*, 2024.

[50] Csaba Szepesvári. Lecture notes in theoretical foundations of reinforcement learning: Lecture 23, tabular mdps. https://rltheory.github.io/lecture-notes/online-rl/lec23/, 2023.

[51] Andrea Tirinzoni, Aymen Al-Marjani, and Emilie Kaufmann. Optimistic pac reinforcement learning: the instance-dependent view. In *International Conference on Algorithmic Learning Theory*, 2023.

[52] Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

[53] Andrew J Wagenmaker, Max Simchowitz, and Kevin Jamieson. Beyond no regret: Instance-dependent pac reinforcement learning. In *Conference on Learning Theory*, 2022.

[54] Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Represent to control partially observed systems: Representation learning with provable sample efficiency. In *International Conference on Learning Representations*, 2023.

[55] Yuanhao Wang, Qinghua Liu, and Chi Jin. Is rlhf more difficult than standard rl? *arXiv preprint arXiv:2306.14111*, 2023.

[56] Christian Wirth, Riad Akrour, Gerhard Neumann, Johannes Fürnkranz, et al. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46, 2017.

[57] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *arXiv preprint arXiv:2306.01693*, 2023.

[58] Wanqi Xue, Qingpeng Cai, Zhenghai Xue, Shuo Sun, Shuchang Liu, Dong Zheng, Peng Jiang, Kun Gai, and Bo An. Prefrec: Recommender systems with human preferences for reinforcing long-term user engagement. *arXiv preprint arXiv:2212.02779*, 2022.

[59] Wenhao Zhan, Masatoshi Uehara, Wen Sun, and Jason D Lee. Pac reinforcement learning for predictive state representations. In *International Conference on Learning Representations*, 2022.

[60] Wenhao Zhan, Masatoshi Uehara, Wen Sun, and Jason D Lee. How to query human feedback efficiently in rl? *arXiv preprint arXiv:2305.18505*, 2023.

[61] Han Zhong, Wei Xiong, Sirui Zheng, Liwei Wang, Zhaoran Wang, Zhuoran Yang, and Tong Zhang. Gec: A unified framework for interactive decision making in mdp, pomdp, and beyond. *arXiv preprint arXiv:2211.01962*, 2023.

[62] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Lemmas and Discussion

A.1 Relation between V_w and V_g

Lemma 1 (Replacing w with g)**.** *For any history-dependent policy π that selects an action $a_h \sim \pi(\tau[h-1], s_h)$, $V_w(\mathbb{M}, \pi) = V_g(\mathbb{M}, \pi)$ holds for any $\mathbb{M}$.*

Proof. By a slight abuse of notation, the following chain of equalities holds. Here, (i) holds since $r_h(s_h, u_h, a_h)$ is a function of s_h, u_h, a_h. Equation (ii) holds since we have already conditioned on $\tau[h]$, which includes s_h, a_h. Equation (iii) holds by the definition of $g_h(\tau[h])$ as the conditional distribution of u_h given $\tau[h]$.

$$
\begin{aligned}
V_w(\mathbb{M}, \pi) &= \mathbb{E}_{\tau^u \sim \mathbb{P}^{w,\pi}} \left[\sum_{h \in \mathcal{H}_p} r_h(s_h, u_h, a_h) \right] \\
&= \sum_{h \in \mathcal{H}_p} \mathbb{E}_{\tau^u \sim \mathbb{P}^{w,\pi}} \left[r_h(s_h, u_h, a_h) \right] \\
&= \sum_{h \in \mathcal{H}_p} \mathbb{E}_{\tau^u[h] \sim \mathbb{P}^{w,\pi}} \left[r_h(s_h, u_h, a_h) \right] \\
&= \sum_{h \in \mathcal{H}_p} \mathbb{E}_{\tau[h] \sim \mathbb{P}^{\pi}} \left[\mathbb{E}_{\tau^u[h] \sim \mathbb{P}^{w,\pi}} \left[r_h(s_h, u_h, a_h) \right] \middle| \tau[h] \right] \\
&\overset{(i)}{=} \sum_{h \in \mathcal{H}_p} \mathbb{E}_{\tau[h] \sim \mathbb{P}^{\pi}} \left[\mathbb{E}_{u_h, s_h, a_h \sim \mathbb{P}^{w,\pi}} \left[r_h(s_h, u_h, a_h) \right] \middle| \tau[h] \right] \\
&\overset{(ii)}{=} \sum_{h \in \mathcal{H}_p} \mathbb{E}_{\tau[h] \sim \mathbb{P}^{\pi}} \left[\mathbb{E}_{u_h \sim \mathbb{P}^{w,\pi}} \left[r_h(s_h, u_h, a_h) \right] \middle| \tau[h] \right] \\
&\overset{(iii)}{=} \sum_{h \in \mathcal{H}_p} \mathbb{E}_{\tau[h] \sim \mathbb{P}^{\pi}} \left[\mathbb{E}_{u_h \sim g_h(\tau[h])} \left[r_h(s_h, u_h, a_h) \right] \right] \\
&= \sum_{h \in \mathcal{H}_p} \mathbb{E}_{\tau \sim \mathbb{P}^{\pi}} \left[\mathbb{E}_{u_h \sim g_h(\tau[h])} \left[r_h(s_h, u_h, a_h) \right] \right] \\
&= \mathbb{E}_{\tau \sim \mathbb{P}^{\pi}} \left[\sum_{h \in \mathcal{H}_p} \mathbb{E}_{u_h \sim g_h(\tau[h])} \left[r_h(s_h, u_h, a_h) \right] \right] \\
&= V_g(\mathbb{M}, \pi)
\end{aligned}
$$

□

A.2 Ignoring Internal Reward-States is Bad for Alignment

We define traditional RL methods as those that output possibly time-dependent Markovian policies. In this section, we provide a toy example showing that there is a PORMDP with good sublinear regret guarantees where any time-dependent Markovian policy has value bounded away from the maximum value. This means that traditional RL methods will always incur linear regret. We hope that this illustrates that RL methods that ignore internal reward-states can be bad for alignment.

Lemma 3 (Markovian policies are not enough)**.** *There is a* PORMDP *where POR-UCRL and POR-UCBVI achieve $poly(H, S, A)\sqrt{T}$ regret, but any Markovian policy is at least $\frac{1}{4}$-suboptimal and so any method that outputs Markovian (possibly time-dependent) policies will lead to linear regret.*

Proof. Consider a PORMDP $\mathbb{M}$ in the setting of [11] (see point (ii) below Definition 2) and set $\mathcal{S} = \{s_1, s_2\}$, $\mathcal{A} = \{a_1, a_2\}$ and $\mathbf{w} = 1 \in \mathbb{R}$. Let the transition matrix be $\mathbb{P}(s' \mid s, a) = \frac{1}{2}$ for all s, a, s'. Let the starting state always be s_1.

Consider the set $\mathcal{T}$ of all trajectories that have a_2 until s_2 appears, and then only have a_1. Choose $\phi(\tau) = \mathbb{1}(\tau \in \mathcal{T})$. The best non-Markovian policy $\pi_\star$ can follow this rule and achieve $\phi(\tau) = 1$ for all $\tau \sim \mathbb{P}^{\pi_\star}$. Thus, $\max_{\pi \in \Pi} V(\mathbb{M}, \pi) = 1$, where Π is given by all history-dependent policies.

On the other hand, consider a Markovian but potentially time-dependent policy π. If $\pi(a_2) = 0$, then its value is zero. If $\pi(a_2) > 0$, then conditioned on the event that s_2 appears first at timestep 1, the expected total reward is at most $\pi_1(a_2)(1 - \pi_2(a_2))$. Conditioned on the event that s_2 appears first at timestep 2, the expected total reward is at most $\pi_1(a_2)\pi_2(a_2)$. Conditioned on seeing s_2 at or after $h = 3$, the expected total reward is certainly at most 1. Using these crude inequalities, we can bound the expected reward of a Markovian policy π by

$$\frac{\pi_1(a_2)(1-\pi_2(a_2))}{2} + \frac{\pi_1(a_2)\pi_2(a_2)}{4} + \sum_{h=3}^{H} \frac{1}{2^h} \leqslant \frac{\pi_1(a_1)(2-\pi_2(a_2))}{4} + \frac{1}{4} \leqslant \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$$

This means that the value of any time-dependent Markovian policy is at most $\frac{3}{4}$ and so any time-dependent Markovian policy is at least $\frac{1}{4}$-suboptimal and incurs $\frac{T}{4}$ regret. □

Recall that we defined traditional RL algorithms as those that output (possibly time-dependent) Markovian policies. Clearly, any traditional RL algorithm in this sense will have at least $\frac{T}{4}$ regret, which is linear regret.

A.3 The Naive Reduction from Dueling to Cardinal PORRL Fails

Lemma 4 (Naive Reduction Lower Bound)**.** *Using any algorithm for cardinal PORRL with sublinear cardinal regret on $\overline{\mathbb{M}}$ with policy class $\Pi' := \Pi \times \Pi$ to get a sequence $(\pi_{1,1}, \pi_{2,1}), \ldots, (\pi_{1,T}, \pi_{2,T})$ leads to linear dueling regret for $\mathbb{M}$ whenever all policies π do not have the same value $V(\mathbb{M}, \pi)$.*

Proof. Define $\pi_\star := \arg\max_{\pi\in\Pi} V(\mathbb{M}, \pi)$ and let $\pi_{\min} := \arg\min_{\pi\in\Pi} V(\mathbb{M}, \pi)$. Then note that

$$\begin{aligned}
\max_{\pi,\pi'\in\Pi} V_D(\mathbb{M}, \pi, \pi') &= \max_{\pi,\pi'\in\Pi} V(\mathbb{M}, \pi) - V(\mathbb{M}, \pi') \\
&= \max_{\pi\in\Pi} V(\mathbb{M}, \pi) + \max_{\pi'\in\Pi} \left[-V(\mathbb{M}, \pi')\right] \\
&= \max_{\pi\in\Pi} V(\mathbb{M}, \pi) - \min_{\pi'\in\Pi} V(\mathbb{M}, \pi') \\
&= V(\mathbb{M}, \pi_\star) - V(\mathbb{M}, \pi_{min})
\end{aligned}$$

Under the naive reduction described in Section 4, a cardinal PORRL algorithm is used to maximize *dueling feedback*. If the algorithm has sublinear cardinal regret, then it will produce duels $(\pi_{1,t}, \pi_{2,t}), t = 1 \to T$, satisfying

$$\sum_{t=1}^{T} \max_{\pi,\pi'\in\Pi} V_D(\mathbb{M}, \pi, \pi') - V_D(\mathbb{M}, \pi_{1,t}, \pi_{2,t}) = o(T)$$

From above, this means that

$$\sum_{t=1}^{T} \left[V(\mathbb{M}, \pi_\star) - V(\mathbb{M}, \pi_{1,t})\right] + \left[V(\mathbb{M}, \pi_{2,t}) - V(\mathbb{M}, \pi_{min})\right] = o(T)$$

Now note that by definition of $\pi_\star$ and π_{min}, both terms are positive. This is the key point. We thus have

$$\sum_{t=1}^{T} V(\mathbb{M}, \pi_\star) - V(\mathbb{M}, \pi_{1,t}) = o(T)$$

$$\sum_{t=1}^{T} V(\mathbb{M}, \pi_{2,t}) - V(\mathbb{M}, \pi_{min}) = o(T)$$

This means that for dueling regret $\mathrm{Regret}_D(T)$, we have the following.

$$\begin{aligned}
\mathrm{Regret}_D(T) &= \sum_{t=1}^{T} \left[V(\mathbb{M}, \pi_\star) - V(\mathbb{M}, \pi_{1,t})\right] + \left[V(\mathbb{M}, \pi_\star) - V(\mathbb{M}, \pi_{2,t})\right] \\
&= \sum_{t=1}^{T} \left[V(\mathbb{M}, \pi_\star) - V(\mathbb{M}, \pi_{1,t})\right] + \left[V(\mathbb{M}, \pi_\star) - V(\mathbb{M}, \pi_{min})\right]
\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{t=1}^{T} \left[V(\mathbb{M}, \pi_{min}) - V(\mathbb{M}, \pi_{2,t}) \right] \\
&= o(T) + T \left[V(\mathbb{M}, \pi_\star) - V(\mathbb{M}, \pi_{min}) \right] \\
&= \Theta(T)
\end{aligned}
$$

Where the last line holds since all policies π do not have the same value $V(\mathbb{M}, \pi)$, and so $V(\mathbb{M}, \pi_\star) - V(\mathbb{M}, \pi_{min}) > 0$. $\square$

B Regret-to-PAC Conversion

When learning in MDPs, we can turn any guarantee on the regret into a corresponding PAC guarantee, the so-called "regret-to-PAC conversion" [30, 39, 53, 51]. Similarly, we want to convert guarantees on the cardinal and dueling regret (see Section 2) into corresponding PAC guarantees, which are more adherent to an offline setting. We provide distinct results for the cardinal and dueling regret below.

Lemma 5 (Cardinal regret to PAC)**.** *For $T \in \mathbb{N}$ and $\delta \in [0,1]$, let* ALG *be an algorithm for cardinal PORRL producing a sequence of policies $(\pi_t)_{t\in[T]}$ with cardinal regret bounded with probability at least $1-\delta$ as*

$$\sum_{t=1}^{T} V(\mathbb{M},\pi_\star) - V(\mathbb{M},\pi_t) \leqslant R(T,\delta) \in \mathbb{R}.$$

Then, a policy $\widehat{\pi}_T \sim \pi_1,\dots,\pi_T$ sampled uniformly satisfies with probability at least $1-2\delta$

$$V(\mathbb{M},\pi_\star) - V(\mathbb{M},\widehat{\pi}_T) \leqslant \frac{R(T,\delta)}{T} + 8Bp\sqrt{\frac{\log(1/\delta)}{T}}.$$

Proof. We consider the sequence of random variables $Y_t = V(\mathbb{M},\pi_\star) - V(\mathbb{M},\pi_t)\ \forall t \in [T]$. Through the Hoeffding's inequality on Y_t and $|r_h| \leqslant B$ we have

$$\begin{aligned} V(\mathbb{M},\pi_\star) - V(\mathbb{M},\widehat{\pi}_T) &= \mathbb{E}[V(\mathbb{M},\pi_\star) - V(\mathbb{M},\pi_t)] \\ &\leqslant \frac{1}{T}\sum_{t=1}^{T}\Big(V(\mathbb{M},\pi_\star) - V(\mathbb{M},\pi_t)\Big) + 8Bp\sqrt{\frac{\log(1/\delta)}{T}} \end{aligned}$$

with probability at least $1-\delta$. Then, combining the latter inequality with the upper bound on the regret and a union bound, we get

$$V(\mathbb{M},\pi_\star) - V(\mathbb{M},\widehat{\pi}_T) \leqslant \frac{R(T,\delta)}{T} + 8Bp\sqrt{\frac{\log(1/\delta)}{T}}$$

with probability at least $1-2\delta$. □

The latter result implies a PAC guarantee of the form $\mathbb{P}(V(\mathbb{M},\pi_\star) - V(\mathbb{M},\widehat{\pi}_T) \geqslant \varepsilon) \leqslant \delta$ for some $\varepsilon > 0$ and $\delta \in [0,1]$ with a number of episodes of order $\widetilde{O}(1/\varepsilon^2)$. An analogous result can be stated for the dueling setting.

Lemma 6 (Dueling regret to PAC)**.** *For $T \in \mathbb{N}$ and $\delta \in [0,1]$, let* ALG *be an algorithm for dueling PORRL producing a sequence of policy pairs $(\pi_{1,t},\pi_{2,t})_{t\in[T]}$ with dueling regret bounded with probability at least $1-\delta$ as*

$$\sum_{t=1}^{T} V(\mathbb{M},\pi_\star) - \frac{V(\mathbb{M},\pi_{1,t}) + V(\mathbb{M},\pi_{2,t})}{2} \leqslant R_D(T,\delta) \in \mathbb{R}.$$

Then, a policy $\widehat{\pi}_T \sim \pi_1,\dots,\pi_T$ sampled uniformly satisfies with probability at least $1-4\delta$

$$V(\mathbb{M},\pi_\star) - V(\mathbb{M},\widehat{\pi}_T) \leqslant \frac{R_D(T,\delta)}{T} + 16Bp\sqrt{\frac{\log(1/\delta)}{T}}.$$

Proof. The proof proceeds as in the previous lemma by applying Hoeffding separately on the sequences $Y_{1,t} = V(\mathbb{M},\pi_\star) - V(\mathbb{M},\pi_{1,t})$ and $Y_{2,t} = V(\mathbb{M},\pi_\star) - V(\mathbb{M},\pi_{2,t})$, then applying a union bound. □

C Proofs for General Optimistic Algorithms for Cardinal PORRL

C.1 Generic Model-Based Optimism using Confidence-Sets

We present a template to get regret bounds for a *generic model-based optimistic algorithm using confidence sets*, which we will later instantiate into POR-UCRL and also use in our reduction from the dueling PORRL to optimistic algorithms for cardinal PORRL.

A generic algorithm using confidence sets is determined by confidence sets $\mathcal{C}_{\mathcal{M}}(\mathcal{D}, \delta)$ based on a dataset $\mathcal{D}$. Maintaining a running dataset $\mathcal{D}_t$, at each step t, we run π_t given by

$$\pi_t, \widetilde{\mathbb{M}}_t := \operatorname*{arg\,max}_{\pi \in \Pi, \mathbb{M} \in \mathcal{C}_{\mathcal{M}}(\mathcal{D}_t, \delta)} V(\mathbb{P}, f, \pi)$$

We obtain a trajectory $\tau_t \sim \mathbb{P}_\star^{\pi_t}$ and append it to $\mathcal{D}_t$ to get $\mathcal{D}_{t+1}$, recompute confidence sets $\mathcal{C}_{\mathcal{M}}(\mathcal{D}_{t+1}, \delta)$, and continue. This algorithm is formally presented in Appendix C.1 below.

Algorithm 2 Generic Confidence-Set Optimism

1: **Input** Known family of reward functions $\{\mathcal{R}_h\}_{h=1}^H$, known model class $\mathcal{M}$ induced by known probability transition kernel class $\mathcal{P}$ and known decoder-induced function class $\mathcal{F}$, confidence level δ.
2: **Initialize** dataset $\mathcal{D}_1 \leftarrow \{\}$ and $\mathcal{C}_{\mathcal{M}}(\mathcal{D}_1, \delta) \leftarrow \mathcal{M}$.
3: **for** $t = 1, ..., T$ **do**
4: **Compute** the optimistic history dependent policy,

$$\pi_t, \widetilde{\mathbb{M}}_t = \operatorname*{arg\,max}_{\pi,\ \mathbb{M} \in \mathcal{C}_{\mathcal{M}}(\mathcal{D}_t, \delta)} V(\mathbb{M}, \pi)$$

5: **Observe** trajectory $\tau_t = \{(s_h^t, a_h^t)\}_{h=1}^H$ and feedback $\{o_h\}_{h \in \mathcal{H}_p}$.
6: **Update** $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup \{\tau_t\}$ and compute new confidence set $\mathcal{C}_{\mathcal{M}}(\mathcal{D}_{t+1}, \delta)$.
7: **end for**

We now make the following assumption about our confidence sets. It essentially controls the effect of shrinking confidence sets for $\mathcal{P}$ and $\mathcal{F}$ on the value. Showing this assumption is the core of proving regret bounds for any instantiation of this generic algorithm. We will see later that it is satisfied by the confidence sets for POR-UCRL.

Assumption 3 (Controlling Value Error due to Confidence Sets, Refined Version)**.** For a transition kernel $\mathbb{P}_\star$ and function $f_\star$, consider any sequence of policies π_t and datasets $\mathcal{D}_t$ that contain $\{\tau_i\}_{i=1}^t$ generated under $(\mathbb{P}_\star^{\pi_t}, f^\star)$. We require that $\mathbb{M}_\star \in \mathcal{C}_{\mathcal{M}}(\mathcal{D}_t, \delta)$ for all t with probability $1 - \delta/16$. We require that there exist problem dependent functions $C_P(\mathcal{M}, T, \delta)$ and $C_F(\mathcal{M}, T, \delta)$ so that for arbitrary sequences $(\mathbb{P}_t, f^t) \in \mathcal{C}_{\mathcal{M}}(\mathcal{D}_t, \delta)$, the following hold with probability $1 - \delta/2$ each.

$$\left| \sum_{t=1}^T V(\mathbb{P}_t, f^t, \pi_t) - V(\mathbb{P}_\star, f^t, \pi_t) \right| = \widetilde{\mathcal{O}}(C_P(\mathcal{M}, T, \delta))$$

$$\left| \sum_{t=1}^T V(\mathbb{P}_\star, f^t, \pi_t) - V(\mathbb{P}_\star, f_\star, \pi_t) \right| = \widetilde{\mathcal{O}}(C_F(\mathcal{M}, T, \delta))$$

Theorem 4 (Regret for Confidence-Set Optimism)**.** *Under Assumption 3, any generic optimistic algorithm using confidence sets $\mathcal{C}_{\mathcal{M}}(\mathcal{D}, \delta)$ satisfies the regret bound*

$$\mathrm{Regret}(T) = \widetilde{\mathcal{O}}\left(C_P(\mathcal{M}, T, \delta) + C_F(\mathcal{M}, T, \delta)\right)$$

Proof. Let $\widetilde{\mathbb{M}}_t$ be given by $\widetilde{\mathbb{P}}_t$ and $\widetilde{f}^t$. Note the following inequalities, where (i) holds with probability 1 by the optimistic definition of π_t.

$$\mathrm{Regret}(T) = \sum_{t=1}^T V(\mathbb{P}_\star, f^\star, \pi_\star) - V(\mathbb{P}_\star, f^\star, \pi_t)$$

$$
\begin{aligned}
&\overset{(i)}{\leqslant} \sum_{t=1}^{T} V(\widetilde{\mathbb{P}}_t, \widetilde{f}^t, \pi_t) - V(\mathbb{P}_\star, f^\star, \pi_t) \\
&\leqslant \sum_{t=1}^{T} \underbrace{V(\widetilde{\mathbb{P}}_t, \widetilde{f}^t, \pi_t) - V(\mathbb{P}_\star, \widetilde{f}^t, \pi_t)}_{(I)} + \underbrace{V(\mathbb{P}_\star, \widetilde{f}^t, \pi_t) - V(\mathbb{P}_\star, f^\star, \pi_t)}_{(II)}
\end{aligned}
$$

We now apply Assumption 3 to bound (I) and (II). We can use the assumption since $\mathcal{D}_t$ contains trajectories $\{\tau_i\}_{i=1}^t$ generated by $\mathbb{P}_\star^{\pi_t}$, $\widetilde{f}^t \in \mathcal{C}_{\mathcal{F}}(\mathcal{D}_t, \delta) \subset \mathcal{F}$ and $\widetilde{\mathbb{P}}^t \in \mathcal{C}_{\mathcal{F}}(\mathcal{D}_t, \delta)$. This immediately gives us that with probability $1 - \delta$

$$
\mathrm{Regret}(T) = \widetilde{\mathcal{O}}(C_F(\mathcal{M}, T, \delta) + C_P(\mathcal{M}, T, \delta))
$$

as desired. □

C.2 Generic Model-Based Optimism using Bonuses

We present a template to get regret bounds for a *generic model-based optimistic algorithm using bonuses*, which we will later instantiate into POR-UCBVI and also use in our reduction from the dueling PORRL to optimistic algorithms for cardinal PORRL.

A generic optimistic algorithm using bonuses relies on bonuses $b_{\mathcal{P}}^{\mathcal{D}}(\mathbb{P}, \pi, \delta), b_{\mathcal{F}}^{\mathcal{D}}(\mathbb{P}, \pi, \delta)$ that depend on a policy π, transition kernel $\mathbb{P}$ and dataset $\mathcal{D}$. It also relies on estimates $\hat{\mathbb{P}}_{\mathcal{D}}$ and $\hat{f}_{\mathcal{D}}$ that depend on $\mathcal{D}$. Maintaining a running dataset $\mathcal{D}_t$, at each step t, we run $\pi_t := \arg\max_{\pi \in \Pi} \widetilde{V}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \hat{f}_{\mathcal{D}_t}, \pi)$, where $\widetilde{V}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \hat{f}_{\mathcal{D}_t}, \pi)$ is given by:

$$V(\hat{\mathbb{P}}_{\mathcal{D}_t}, \hat{f}_{\mathcal{D}_t}, \pi) + b_{\mathcal{F}}^{\mathcal{D}_t}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi, \delta) + z(Bp) b_{\mathcal{P}}^{\mathcal{D}_t}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi, \delta)$$

where z is defined below. We obtain a trajectory $\tau_t \sim \mathbb{P}_\star^{\pi_t}$ and append it to $\mathcal{D}_t$ to get $\mathcal{D}_{t+1}$, compute new bonuses and estimates, and continue. This algorithm is formally presented in Appendix C.2.

Algorithm 3 Generic Bonus-Based Optimism

1: **Input** Known family of reward functions $\{\mathcal{R}_h\}_{h=1}^H$, method $\texttt{Est}(\mathcal{D})$ to estimate $\hat{\mathbb{P}}_{\mathcal{D}}$ and $\hat{f}_{\mathcal{D}}$ from dataset $\mathcal{D}$, bonus functions $b_{\mathcal{F}}^{\mathcal{D}}(\mathbb{P}, \pi, \delta)$ and $b_{\mathcal{P}}^{\mathcal{D}}(\mathbb{P}, \pi, \delta)$, confidence level δ
2: **Initialize** $\mathcal{D}_1 \leftarrow \{\}$, initialize $\hat{f}^{\mathcal{D}_1}, \hat{\mathbb{P}}_{\mathcal{D}_1}$ arbitrarily.
3: **for** $t = 1, ..., T$ **do**
4: **Compute** optimistic history dependent policy,

$$\pi_t = \arg\max_{\pi} V(\hat{\mathbb{P}}_{\mathcal{D}_t}, \hat{f}_{\mathcal{D}_t}, \pi) + b_{\mathcal{F}}^{\mathcal{D}_t}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi, \delta) + z(Bp)(b_{\mathcal{P}}^{\mathcal{D}_t}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi, \delta))$$

5: **Observe** trajectory $\tau_t = \{(s_h^t, a_h^t)\}_{h=1}^H$ and feedback $\{o_h\}_{h \in \mathcal{H}_p}$.
6: **Compute** new estimates $\hat{f}^{\mathcal{D}_{t+1}}, \hat{\mathbb{P}}^{\mathcal{D}_{t+1}} \leftarrow \texttt{Est}(\mathcal{D}_{t+1})$ and compute new bonus functions $b_{\mathcal{F}}^{\mathcal{D}_{t+1}}(\hat{f}^{\mathcal{D}_{t+1}}, \cdot, \delta), b_{\mathcal{P}}^{\mathcal{D}_{t+1}}(\hat{\mathbb{P}}^{\mathcal{D}_{t+1}}, \cdot, \delta)$.
7: **end for**

We now make the following assumption about our bonuses. Showing this assumption is the core of proving regret bounds for any instantiation of this generic algorithm. We will see later that it is satisfied by the bonuses for POR-UCBVI.

Assumption 4 (Controlling Value Error via Bonuses)**.** For a transition kernel $\mathbb{P}_\star$ and function $f_\star$, consider any sequence of policies π_t and datasets $\mathcal{D}_t$ that contain $\{\tau_i\}_{i=1}^t$ generated under $(\mathbb{P}_\star^{\pi_t}, f^\star)$. We require that for sequences $\hat{\mathbb{P}}_{\mathcal{D}_t}$ and $\hat{f}_{\mathcal{D}_t}$ and any sequence $f^t \in \mathcal{F}$, the following hold.

- *Bounding effect of error in $\mathcal{F}$:* With probability $1 - \delta/32$, for any $\mathbb{P}$ and uniformly over all policies π, $|V(\mathbb{P}, \hat{f}_{\mathcal{D}_t}, \pi) - V(\mathbb{P}, f^\star, \pi)| \leqslant b_{\mathcal{F}}^{\mathcal{D}_t}(\mathbb{P}, \pi, \delta)$ and there is a function $C_F(\mathcal{M}, T, \delta)$ so that $\sum_{t=1}^T b_{\mathcal{F}}^{\mathcal{D}_t}(\mathbb{P}_\star, \pi_t, \delta) = \widetilde{\mathcal{O}}(C_F(\mathcal{M}, T, \delta))$ with probability $1 - \delta/32$
- *Bounding effect of error in $\mathcal{P}$:* For any function $\mu : \Gamma_H \to \mathbb{R}$ bounded by D, there is a function $z(D) \geqslant D$ so that the following holds uniformly over all policies π with probability $1 - \delta/32$.

$$\mathbb{E}_{\tau \sim (\hat{\mathbb{P}}_{\mathcal{D}_t})^\pi} \mu(\tau) - \mathbb{E}_{\tau \sim \mathbb{P}_\star^\pi} \mu(\tau) \leqslant z(D) b_{\mathcal{P}}^{\mathcal{D}_t}(\mathbb{P}_\star, \pi, \delta)$$

  The statement also holds if we switch $\mathbb{P}_\star$ and $\hat{\mathbb{P}}_{\mathcal{D}_t}$. Additionally, the statement holds for a suitable D if we replace $\mathbb{E}_{\tau \sim \mathbb{P}^\pi} \mu(\tau)$ with $b_{\mathcal{P}}(\mathbb{P}, \pi, \delta)$ or $b_{\mathcal{F}}(\mathbb{P}, \pi, \delta)$.[10] Finally, there is a function $C_P(\mathcal{M}, T, \delta)$ so that $\sum_{t=1}^T b_{\mathcal{P}}^{\mathcal{D}_t}(\mathbb{P}_\star, \pi_t, \delta) = \widetilde{\mathcal{O}}(C_P(\mathcal{M}, T, \delta))$ with probability $1 - \delta/32$.

Theorem 5 (Regret for Bonus-Based Optimism)**.** *Under Assumption 4, with $z_1(D) = z(D) + z(2D) + z(2z(D))$, any generic optimistic algorithm using bonuses satisfies*

$$\text{Regret}(T) = \widetilde{\mathcal{O}}\left(C_F(\mathcal{M}, T, \delta) + z_1(Bp) C_P(\mathcal{M}, T, \delta)\right)$$

[10] This would instantly hold with $D = Bp$ if $b_{\mathcal{F}}(\mathbb{P}, \pi, \delta) := \mathbb{E}_{\tau \sim \mathbb{P}^\pi} b_{\mathcal{F}}(\tau, \delta)$ for some trajectory level bonus $b_{\mathcal{F}}(\tau, \delta)$, and similarly for $\mathcal{P}$.

Proof. Note that we can use Assumption 4 since $\mathcal{D}_t$ contains trajectories $\{\tau_i\}_{i=1}^t$ generated by $\mathbb{P}_\star^{\pi_t}$, $\hat{f}_{\mathcal{D}_t} \in \mathcal{F}$ is computed using $\mathcal{D}_t$ and $\hat{\mathbb{P}}_{\mathcal{D}_t}$ is computed using $\mathcal{D}_t$. Also note that WLOG, $b_{\mathcal{P}}^{\mathcal{D}_t}(\mathbb{P}, \pi, \delta) \leqslant 2$ always holds since we can otherwise clip it at 2 and our assumption will still hold. Similarly, WLOG $b_{\mathcal{F}}^{\mathcal{D}_t}(\mathbb{P}, \pi, \delta) \leqslant 2z(Bp)$, otherwise we can clip it at 1 and our assumption will still hold. Now note the following inequalities.

$$
\begin{aligned}
\mathrm{Regret}(T) &= \sum_{t=1}^T V(\mathbb{P}_\star, f^\star, \pi_\star) - V(\mathbb{P}_\star, f^\star, \pi_t) \\
&\overset{(i)}{\leqslant} \sum_{t=1}^T V(\hat{\mathbb{P}}_{\mathcal{D}_t}, f^\star, \pi_\star) + z(Bp)(b_{\mathcal{P}}^{\mathcal{D}_t}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_\star, \delta)) - V(\mathbb{P}_\star, f^\star, \pi_t) \\
&\overset{(ii)}{\leqslant} \sum_{t=1}^T V(\hat{\mathbb{P}}_{\mathcal{D}_t}, \hat{f}_{\mathcal{D}_t}, \pi_\star) + b_{\mathcal{F}}^{\mathcal{D}_t}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_\star, \delta) + z(Bp)(b_{\mathcal{P}}^{\mathcal{D}_t}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_\star, \delta)) - V(\mathbb{P}_\star, f^\star, \pi_t) \\
&\overset{(iii)}{\leqslant} \sum_{t=1}^T V(\hat{\mathbb{P}}_{\mathcal{D}_t}, \hat{f}_{\mathcal{D}_t}, \pi_t) + b_{\mathcal{F}}^{\mathcal{D}_t}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_t, \delta) + z(Bp)(b_{\mathcal{P}}^{\mathcal{D}_t}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_t, \delta)) - V(\mathbb{P}_\star, f^\star, \pi_t) \\
&= \sum_{t=1}^T V(\hat{\mathbb{P}}_{\mathcal{D}_t}, \hat{f}_{\mathcal{D}_t}, \pi_\star) - V(\mathbb{P}_\star, f^\star, \pi_t) + b_{\mathcal{F}}^{\mathcal{D}_t}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_t, \delta) + z(Bp)(b_{\mathcal{P}}^{\mathcal{D}_t}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_t, \delta)) \\
&= \sum_{t=1}^T V(\hat{\mathbb{P}}_{\mathcal{D}_t}, \hat{f}_{\mathcal{D}_t}, \pi_\star) - V(\hat{\mathbb{P}}_{\mathcal{D}_t}, f^\star, \pi_\star) + V(\hat{\mathbb{P}}_{\mathcal{D}_t}, f^\star, \pi_\star) - V(\mathbb{P}_\star, f^\star, \pi_t) \\
&\qquad + \sum_{t=1}^T b_{\mathcal{F}}^{\mathcal{D}_t}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_t, \delta) + z(Bp)(b_{\mathcal{P}}^{\mathcal{D}_t}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_t, \delta))
\end{aligned}
$$

Here, inequality (i) holds with probability $1 - \delta/16$ by the second point in Assumption 4. Inequality (ii) holds with probability $1 - \delta/16$ by the first point in Assumption 4. Inequality (iii) holds with probability 1 by the optimistic definition of π_t. Continuing, we have

$$
\begin{aligned}
\mathrm{Regret}(T) &\overset{(iv)}{\leqslant} 2\sum_{t=1}^T b_{\mathcal{F}}^{\mathcal{D}_t}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_t, \delta) + z(Bp)(b_{\mathcal{P}}^{\mathcal{D}_t}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_t, \delta)) \\
&\overset{(v)}{\leqslant} 2\sum_{t=1}^T b_{\mathcal{F}}^{\mathcal{D}_t}(\mathbb{P}_\star, \pi_t, \delta) + 2z(Bp)(b_{\mathcal{P}}^{\mathcal{D}_t}(\mathbb{P}_\star, \pi_t, \delta)) + z(Bp)(b_{\mathcal{P}}^{\mathcal{D}_t}(\mathbb{P}_\star, \pi_t, \delta)) \\
&\qquad + 2z(z(Bp))(b_{\mathcal{P}}^{\mathcal{D}_t}(\mathbb{P}_\star, \pi_t, \delta)) \\
&= \mathcal{O}\left(\sum_{t=1}^T b_{\mathcal{F}}^{\mathcal{D}_t}(\mathbb{P}_\star, \pi_t, \delta) + z_1(Bp)(b_{\mathcal{P}}^{\mathcal{D}_t}(\mathbb{P}_\star, \pi_t, \delta))\right)
\end{aligned}
$$

Here, inequality (iv) holds with probability $1 - \delta/8$ by a union bound over the first and the second point in Assumption 4. Finally, inequality (v) holds with probability $1 - \delta/4$ by a union bound over four applications of the second point of Assumption 4. Finally, we use a union bound over both points of Assumption 4 to conclude that with probability $1 - \delta/8$

$$
\mathcal{O}\left(\sum_{t=1}^T b_{\mathcal{F}}^{\mathcal{D}_t}(\mathbb{P}_\star, \pi_t, \delta) + z_1(Bp)(b_{\mathcal{P}}^{\mathcal{D}_t}(\mathbb{P}_\star, \pi_t, \delta))\right) = \widetilde{\mathcal{O}}\left(C_F(\mathbb{M}, T, \delta) + z_1(Bp)C_P(\mathbb{M}, T, \delta)\right)
$$

By taking a union bound over the events of all inequalities above, we have that with probability $1 - \delta$

$$
\mathrm{Regret}(T) = \widetilde{\mathcal{O}}\left(C_F(\mathbb{M}, T, \delta) + z_1(Bp)C_P(\mathbb{M}, T, \delta)\right)
$$

as desired. □

C.3 Generic Model-Free Optimism

Algorithm 4 Generic Model-Free Optimism

Input Known Bellman-complete class of Q-functions $\mathcal{Q}$, confidence level δ.
Initialize dataset $\mathcal{D}_1 \leftarrow \{\}$ and $\mathcal{C}_{\mathcal{Q}}(\mathcal{D}_1, \delta) \leftarrow \mathcal{Q}$.
for $t = 1, ..., T$ **do**
  $\tau[0] \leftarrow ()$
  for $h = 1, \ldots H$ **do**
    Play $a_h^t, Q_h^t \leftarrow \arg\max_{a, Q \in \mathcal{C}_{\mathcal{Q}}(\mathcal{D}_t, \delta)} Q_h(\tau[h], a)$ and observe feedback o_h^t
  end for
  Update $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup \{\tau, (o_1^t, \ldots o_H^t\}$
  Compute $\mathcal{C}_{\mathcal{Q}}(\mathcal{D}_{t+1}, \delta)$
end for

Note that the method for choosing actions a_h^t at time t induces a history dependent policy π_t, whose suboptimality is what we use to define regret. Regret is still given by

$$\text{Regret}(T) = \sum_{t=1}^{T} V(\mathbb{M}_\star, \pi_\star) - V(\mathbb{M}_\star, \pi_t)$$

We now make the following assumption about our confidence sets. Showing this assumption is the core of proving regret bounds for any instantiation of this generic algorithm. We know that this is satisfied by GOLF using the BE-dimension. We will show that in our case, it is also satisfied by a more refined notion known as the α-HABE dimension (the α-history aware Bellman eluder dimension).

Assumption 5. For a Q-function $Q^\star$ induced by model $\mathbb{M}_\star$, consider any sequence of policies π_t and datasets $\mathcal{D}_t$ that contain $\{\tau_i\}_{i=1}^t$ generated under $\mathbb{M}_\star$. We require that $Q^\star \in \mathcal{C}_{\mathcal{Q}}(\mathcal{D}_t, \delta)$ for all t with probability $1 - \delta/16$. We require that there exists a problem dependent function $C_Q(\mathcal{Q}, T, \delta)$, so that for arbitrary sequences $Q^t \in \mathcal{C}_{\mathcal{Q}}(\mathcal{D}_t, \delta)$, the following holds for all h with probability $1 - \delta/2$.

$$\sum_{j=1}^{t} |\mathbb{E}_{\mu_h(Q^t)}[Q_h^t - \mathcal{T}_h Q_h^{j+1}]| \leqslant C_Q(\mathcal{Q}, T, \delta)$$

Theorem 6 (Regret for Generic Model-Free Optimism)**.** *If the confidence sets $\mathcal{C}_{\mathcal{Q}}(\mathcal{D}, \delta)$ used in Algorithm 4 satisfy Assumption 5, then the regret of Algorithm 4 is bounded by*

$$\text{Regret}(T) = O\left(H C_Q(\mathcal{Q}, T, \delta)\right)$$

Proof. Note that $V(\mathbb{M}_\star, \pi_\star) = \max_a Q_1^\star(s_1, a) \leqslant \max_a Q_1^t(s_1, a)$ for all t, giving us the following result by the policy loss decomposition in [29].

$$\begin{aligned}
\text{Regret}(T) &= \sum_{t=1}^{T} V(\mathbb{M}_\star, \pi_\star) - V(\mathbb{M}_\star, \pi_t) \\
&\leqslant \sum_{t=1}^{T} \max_a Q_1^t(s_1, a) - V(\mathbb{M}_\star, \pi_t) \\
&= \sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E}_{\mu_h(Q^t)}[Q_h^t - \mathcal{T}_h Q_h^{j+1}] \\
&= \sum_{h=1}^{H} \sum_{t=1}^{T} \mathbb{E}_{\mu_h(Q^t)}[Q_h^t - \mathcal{T}_h Q_h^{j+1}] \\
&= O(H C_Q(\mathcal{Q}, T, \delta))
\end{aligned}$$

where the last line holds by Assumption 5. □

D Details and Proofs for Cardinal POR-UCRL

We now instantiate Algorithm 2 using standard confidence sets to get POR-UCRL. We show that they satisfy Assumption 3 and get regret bounds for the algorithm. Note that our algorithm is **crucially different** from naively summarizing the history to define a modified state space, since we are separating the use of history summarization for getting confidence sets f from using only the current state while learning the Markovian transitions $\mathbb{P}$. In this case, it is a priori unclear if we can use ideas from optimism to prove guarantees with a favorable (non-exponential) dependence on the complexity of transitions.

Recall that given a dataset of the first t trajectory samples $\{\tau_i\}_{i=1}^t$ and an index $h \in [H]$, we consider the following least squares objective to estimate f:

$$\widehat{f}_h^{t+1} = \arg\min_{f_h \in \mathcal{F}_h} \sum_{i=1}^{t} \left(\sigma_h(f_h(\tau_i[h])) - o_h^i\right)^2$$

Simple least squares guarantees imply the lemma below.

Lemma 7 (Concentration for $\sigma \circ f_h$)**.** *Define*

$$\mathrm{MSE}_{h,t}(f_h, f_h') := \sum_{i=1}^{t} \left(\sigma_h(f_h(\tau_i[h])) - \sigma_h(f_h'(\tau_i[h]))\right)^2$$

Also define $\bar{\beta}_{h,t}(\delta) = \eta_h^2 \log\left(\frac{N\left(\mathcal{F}_h, \frac{B}{T}, \|\cdot\|_\infty\right)}{\delta}\right) + \alpha_{h,t}$ *with* $\alpha_{h,t} := \frac{tB + t\eta_h \log\left(\frac{t}{\delta}\right)}{T}$*. Then* $f_h^\star$ *simultaneously satisfies* $\mathrm{MSE}_{h,t}(f_h^\star, \widehat{f}_h^{(t+1)}) \leqslant \bar{\beta}_{h,t}\left(\frac{\delta}{2t^2 H}\right)$ *for all* h, t *with probability* $1 - \delta/32$*.*

Proof. We apply Lemma 6 in [10] and the last statement in its proof to each h separately with the function class in the lemma set to $\{\sigma_h \circ f_h | f_h \in \mathcal{F}_h\}$, $P = 1$, $\mathbf{x}_{t,p} = \mathbf{x}_{t,1} = \tau_t[h]$ and misspecification $\varepsilon = 0$ (decoupled from the Eluder dimension's ε). We also note that o_h^t are η_h-subgaussian samples with mean $\sigma_h(f_h(\tau[h]))$. This gives us that each of event indexed by h, t below holds with probability at least $1 - \frac{\delta}{2t^2 H}$.

$$\sum_{i=1}^{t} \left(\sigma_h(f_h(\tau_i[h])) - \sigma_h(f_h'(\tau_i[h]))\right)^2 \leqslant \bar{\beta}_{h,t}\left(\frac{\delta}{2t^2 H}\right)$$

So, the events all simultaneously hold with probability at least $1 - \delta$ by a union bound. □

Recall the definition of our confidence sets below.

Confidence Sets for POR-UCRL. We instantiate the generic optimistic algorithm using confidence sets by defining $\mathcal{C}_{\mathcal{M}}(\mathcal{D}_t, \delta) := \mathcal{C}_{\mathcal{P}}^t(\delta) \times \mathcal{C}_{\mathcal{F}}^t(\delta)$ as our confidence sets below. We name the resulting algorithm POR-UCRL. We use the data from trajectories $\{\tau_i\}_{i=1}^t$ to build the confidence sets $\mathcal{C}_{\mathcal{F}}^{t+1}(\delta) = \prod_h \mathcal{C}_h^{t+1}(\delta)$ with $\mathcal{C}_h^{t+1}(\delta)$ defined below, where $\beta_{h,t}(\delta) := \bar{\beta}_{h,t}\left(\frac{\delta}{2t^2 H}\right)$.

$$\mathcal{C}_h^{t+1}(\delta) := \left\{ f_h \in \mathcal{F}_h \middle| \mathrm{MSE}_{h,t}(f_h^\star, \widehat{f}_h^{(t+1)}) \leqslant \beta_{h,t}(\delta) \right\}$$

We also use the MLE estimate for $\mathbb{P}$ after t episodes to define $\hat{\mathbb{P}}^t(\cdot \mid s, a) := \frac{N_t(s,a,s')}{N_t(s,a)}$. Now for $\zeta(n, \delta) = 2\sqrt{\frac{S\log(2) + \log(n(n+1)SA/\delta)}{2n}}$, define $\mathcal{C}_{\mathbb{P}}^t(\delta)$ as below:

$$\left\{ \mathbb{P} \middle| \|\mathbb{P}(\cdot \mid s, a) - \hat{\mathbb{P}}_t(\cdot \mid s, a)\|_1 \leqslant \zeta(N_t(s, a), \delta) \forall s, a \right\}$$

Confidence Sets for POR-UCRL in case $\mathbb{P}_\star$ is known. For known-model UCRL, the confidence sets $\mathcal{C}_{\mathcal{F}}^t(\delta)$ are still as above, but $\mathcal{C}_{\mathcal{P}}^t(\delta) := \{\mathbb{P}_\star\}$

For completeness, we repeat the algorithm POR-UCRL here, which is an instantiation of Algorithm 2, the generic optimistic algorithm using confidence sets.

Algorithm 5 POR-UCRL

1: **Input:** Known family of reward functions $\{\mathcal{R}_h\}_{h=1}^H$, known probability transition kernel class $\mathcal{P}$ and known decoder-induced function class $\mathcal{F}$, confidence level δ.
2: **Initialize** dataset $\mathcal{D}_1 \leftarrow \{\}$ and $\mathcal{C}_{\mathcal{F}}(\mathcal{D}_1, \delta) \leftarrow \prod_{h=1}^H \mathcal{F}_h, \mathcal{C}_{\mathcal{P}}(\mathcal{D}_1, \delta) \leftarrow \mathcal{P}$.
3: **for** $t = 1, ..., T$ **do**
4: **Compute** the optimistic history dependent policy,

$$\pi_t, \widetilde{f}_t, \widetilde{\mathcal{P}}_t = \underset{\pi,\ \mathcal{F}\in\mathcal{C}_{\mathcal{F}}(\mathcal{D}_t,\delta), \mathbb{P}\in\mathcal{P}(\mathcal{D}_t,\delta)}{\arg\max} V(\mathbb{P}, f, \pi)$$

5: **Collect** trajectory $\tau_t = \{(s_h^t, a_h^t)\}_{h=1}^H$ and feedback $\{o_h\}_{h\in\mathcal{H}_p}$ by sampling from $\mathbb{P}_\star^{\pi_t}$ with true decoder-induced function $f_\star$.
6: **Update** $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup \{\tau_t\}, \hat{\mathbb{P}}_{t+1}, \hat{f}_h^{t+1}$ for all h
7: **Compute** new confidence sets $\mathcal{C}_{\mathcal{F}}(\mathcal{D}_{t+1}, \delta) \leftarrow \prod_{h=1}^H \mathcal{C}_h^{t+1}(\delta)$ and $\mathcal{C}_{\mathcal{P}}(\mathcal{D}_{t+1}, \delta)$ where

$$\mathcal{C}_h^{t+1}(\delta) \leftarrow \left\{ f_h \in \mathcal{F}_h \middle| \mathrm{MSE}_{h,t}(f_h^\star, \widehat{f}_h^{(t+1)}) \leqslant \beta_{h,t}(\delta) \right\}$$

$$\mathcal{C}_{\mathcal{P}}(\mathcal{D}_{t+1}, \delta) \leftarrow \left\{ \mathbb{P} \middle| \|\mathbb{P}(\cdot \mid s, a) - \hat{\mathbb{P}}_{t+1}(\cdot \mid s, a)\|_1 \leqslant \zeta(N_{t+1}(s, a), \delta) \forall s, a \right\}$$

8: **end for**

We will now show our regret bound.

Theorem 1 (POR-UCRL Regret)**.** *Under Assumption 1, the regret* $\mathrm{Regret}(T)$ *of POR-UCRL is bounded by the following with probability at least* $1 - \delta$

$$\widetilde{\mathcal{O}}\left(\left(pS\sqrt{HA} + \sum\nolimits_{h\in\mathcal{H}_p} \sqrt{d_{E,h} d_{C,h}}\right)\sqrt{T}\right)$$

where $d_{E,h} = \dim_E\left(\mathcal{F}_h, \frac{B}{T}\right)$ *and* $d_{C,h} = \log(\mathcal{N}(\mathcal{F}_h, 1/T, \|\cdot\|_\infty))$.

D.1 Showing that Assumption 3 is Satisfied

D.1.1 Bounding Reward Model Deviations

Lemma 8 (Bounding Reward Model Deviations)**.** *Consider decoder-induced functions* $\{f_h\}_{h\in\mathcal{H}_p}$ *satisfying* $|f_h| \leqslant B$ *that induce value functions* $V(\mathbb{P}, f, \pi)$*. For any sequence of policies* π_t*, if the confidence* $\mathcal{C}_{\mathcal{F}}^t(\delta)$ *is generated using data* $\tau_i \sim \mathbb{P}_\star^{\pi_i}, i = 1 \to t$ *and* $\widetilde{f}^t \in \mathcal{C}_{\mathcal{F}}^t(\delta)$ *is an arbitrary sequence of functions, then we have the following with probability* $1 - \delta/4$*.*

$$\left|\sum_{t=1}^T V(\mathbb{P}_\star, \widetilde{f}^t, \pi_t) - V(\mathbb{P}_\star, f^\star, \pi_t)\right|$$

is bounded by

$$\mathcal{O}\left(Bp\sqrt{T\log(T/\delta)} + \sum_{h\in\mathcal{H}_p} B\kappa_{2,h} d_{E,h} + \sum_{h\in\mathcal{H}_p} \kappa_{2,h}\sqrt{d_{E,h}\beta_{h,T}(\delta)T}\right)$$

Proof.

$$\begin{aligned}
\sum_{t=1}^T V(\mathbb{P}_\star, \widetilde{f}^t, \pi_t) - V(\mathbb{P}_\star, f^\star, \pi_t) &= \sum_{t=1}^T \mathbb{E}_{\tau\sim\mathbb{P}_\star^{\pi_t}}\left[\sum_{h=1}^H \widetilde{f}_h^t(\tau[h])\right] - \mathbb{E}_{\tau\sim\mathbb{P}_\star^{\pi_t}}\left[\sum_{h=1}^H f_h^\star(\tau[h])\right] \\
&= \sum_{t=1}^T \mathbb{E}_{\tau\sim\mathbb{P}_\star^{\pi_t}}\left[\sum_{h\in\mathcal{H}_p} \widetilde{f}_h^t(\tau[h])\right] - \mathbb{E}_{\tau\sim\mathbb{P}_\star^{\pi_t}}\left[\sum_{h\in\mathcal{H}_p} f_h^\star(\tau[h])\right] \\
&= \sum_{t=1}^T \mathbb{E}_{\tau\sim\mathbb{P}_\star^{\pi_t}}\left[\sum_{h\in\mathcal{H}_p} \widetilde{f}_h^t(\tau[h]) - f_h^\star(\tau[h])\right]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{t=1}^{T} \left[\sum_{h \in \mathcal{H}_p} \widetilde{f}_h^t(\tau_t[h]) - f_h^\star(\tau_t[h]) + X_{1,t} + X_{2,t} \right] \\
&\overset{(ii)}{\leqslant} \sum_{t=1}^{T} \left[\sum_{h \in \mathcal{H}_p} \widetilde{f}_h^t(\tau_t[h]) - f_h^\star(\tau_t[h]) \right] + \mathcal{O}\left(Bp\sqrt{T \log(T/\delta)} \right)
\end{aligned}
$$

where

$$
\begin{aligned}
X_{1,t} &:= \mathbb{E}_{\tau \sim \mathbb{P}^\pi} \left[\sum_{h \in \mathcal{H}_p} \widetilde{f}_h^t(\tau[h]) \right] - \left[\sum_{h \in \mathcal{H}_p} \widetilde{f}_h^t(\tau_t[h]) \right] \\
X_{2,t} &:= \left[\sum_{h \in \mathcal{H}_p} f_h^\star(\tau_t[h]) \right] - \mathbb{E}_{\tau \sim \mathbb{P}^\pi} \left[\sum_{h \in \mathcal{H}_p} f_h^\star(\tau[h]) \right]
\end{aligned}
$$

Inequality (i) follows by the definition of π_t and $\widetilde{f}_h^t$ – that is, by optimism. Inequality (ii) holds with probability at least $1 - \delta$ since $X_{1,t}$ and $X_{2,t}$ are both martingales with respect to the filtration $\mathcal{G}_t$ given by the data of trajectories $\{\tau_s\}_{s=1}^{t-1}$. Also, $|X_{1,t}|, |X_{2,t}| \leqslant Bp$. We can thus apply the Azuma-Hoeffding inequality twice to obtain inequality (ii).

Continuing, note the following.

$$
\begin{aligned}
&\sum_{t=1}^{T} V(\mathbb{P}_\star, \widetilde{f}^t, \pi_t) - V(\mathbb{P}_\star, f^\star, \pi_t) \\
&\leqslant \left[\sum_{h=1}^{H} f_h^t(\tau_t[h]) - f_h^\star(\tau_t[h]) \right] + Bp\sqrt{T \log(T/\delta)} \\
&\leqslant \left[\sum_{h=1}^{H} \kappa_{2,h}\sigma_h(f_h^t(\tau_t[h])) - \kappa_{2,h}\sigma_h(f_h^\star(\tau_t[h])) \right] + Bp\sqrt{T \log(T/\delta)} \\
&\leqslant \kappa_{2,h} \sum_{t=1}^{T} \sum_{h \in \mathcal{H}_p} \underbrace{\max_{f_h, f_h' \in \mathcal{W}_h^t(\delta)} \sigma_h(f_h(\tau_t[h])) - \sigma_h(f_h'(\tau_t[h]))}_{=:\bar{\gamma}_{h,t}(\tau_t[h],\delta)} + Bp\sqrt{T \log(T/\delta)} \\
&= \kappa_{2,h} \sum_{h \in \mathcal{H}_p} \left[\sum_{t=1}^{T} \bar{\gamma}_{h,t}(\tau_t[h], \delta) \right] + Bp\sqrt{T \log(T/\delta)}
\end{aligned}
$$

The sum of these maximum uncertainty evaluations can be upper bounded using the Eluder dimension. The inequality below holds by applying Lemma 3 in [10] for each h separately, with the function class in the lemma set to $\{\sigma_h \circ f_h | f_h \in \mathcal{F}_h\}$, $P = 1$, $\mathbf{x}_{t,p} = \mathbf{x}_{t,1} = \tau_t[h]$ and misspecification $\varepsilon = 0$ (decoupled from the Eluder dimension's ε). We also recall that o_h^t are η_h-subgaussian samples with mean $\sigma_h(f_h(\tau_t[h]))$. We obtain

$$
\sum_{t=1}^{T} \bar{\gamma}_{h,t}(\tau_t[h], \delta) \leqslant \mathcal{O}\left(Bd_{E,h} + \sqrt{d_{E,h}\beta_{h,T}(\delta)T} \right)
$$

Where $d_{E,h} = \dim_E\left(\mathcal{F}_h, \frac{B}{T}\right)$ is the Eluder dimension of $\mathcal{F}_h$ and $\beta_{h,T}(\delta) = \bar{\beta}_{h,t}\left(\frac{\delta}{2t^2 H}\right)$. Therefore, we have our result.

$$
\sum_{t=1}^{T} V(\mathbb{P}_\star, \widetilde{f}^t, \pi_t) - V(\mathbb{P}_\star, f^\star, \pi_t)
$$

is bounded by

$$\mathcal{O}\left(\sum_{h\in\mathcal{H}_p} B\kappa_{2,h}d_{E,h} + \sum_{h\in\mathcal{H}_p}\kappa_{2,h}\sqrt{d_{E,h}\beta_{h,T}(\delta)T} + Bp\sqrt{T\log(T/\delta)}\right)$$

Note that this entire argument can be repeated with $f_\star$ and $\widetilde{f}^t$ switched, by the symmetry of the definition of $\bar{\gamma}_{h,t}(\tau_t[h],\delta)$ and the fact that the negative of a martingale is also a martingale. □

D.1.2 Bounding Probability Model Deviations

Lemma 9 (Bounding Probability Model Deviations)**.** *Consider an arbitrary sequence of functions $f^t \in \mathcal{F}$ satisfying $|f_h| \leqslant B$ that induce value functions $V(\mathbb{P}, f, \pi)$. For any sequence of policies π_t, if the confidence $\mathcal{C}^t_{\mathcal{P}}(\delta)$ is generated using data that includes $\tau_i \sim \mathbb{P}^{\pi_i}_\star, i = 1 \to t$ and $\widetilde{\mathbb{P}}_t \in \mathcal{C}^t_{\mathcal{P}}(\delta)$ is an arbitrary sequence of transition structures, then we have the following with probability $1-\delta/4$.*

$$\left|\sum_{t=1}^T V(\widetilde{\mathbb{P}}_t, f^t, \pi_t) - V(\mathbb{P}_\star, f^t, \pi_t)\right| \leqslant \mathcal{O}\left(c_\delta Bp\sqrt{SAHT} + c_\delta BpHSA + Bp\sqrt{pT\log(T/\delta)}\right)$$

where $c_\delta := 8\sqrt{S\log(2) + \log(HTSA/\delta)}$.

Proof. We first show the following.

Lemma 10.

$$\sum_{t=1}^T V(\widetilde{\mathbb{P}}_t, f^t, \pi_t) - V(\mathbb{P}_\star, f^t, \pi_t) \leqslant \sum_{t=1}^T\sum_{h=1}^H 2p\zeta(N_t(s^t_h, a^t_h), \delta) + \mathcal{O}\left(Bp\sqrt{pT\log(T/\delta)}\right)$$

Recall that we denote the Bellman operator by $\mathcal{T}^\pi$ where $\mathcal{T}^\pi f = \mathbb{E}_{a\sim\pi}\mathbb{P}f$. Momentarily define the following for $\tau = (\tau_{l-1}, s_l, \tau')$, where τ_{l-1} is an arbitrary trajectory of length $l-1 \leqslant H$, and s_l is an arbitrary state.

$$\begin{aligned} V^t_{l,\mathbb{P}}(\tau_{l-1}, s_l) &:= \mathbb{E}_{\tau'\sim\mathbb{P}^{\pi_t}}\left[\sum_{h=l}^H \widetilde{f}^t_h(\tau[h])\right] \\ &= \mathbb{E}_{\tau'\sim\mathbb{P}^{\pi_t}}\left[\sum_{h\in\mathcal{H}_p, h\geqslant l} \widetilde{f}^t_h(\tau[h])\right] \end{aligned}$$

So in the definition above, the first $l-1$ observations in τ come from τ_{l-1} while the rest are generated by the input $\mathbb{P}$ starting with state s_l. Note that $V(\mathbb{P}, f^t, \pi_t) = V^t_{1,\mathbb{P}}(\varnothing, s_1)$. Also note that by the Bellman equation, we have the following.

$$\begin{aligned} V^t_{l,\mathbb{P}}(\tau_t[l-1], s_l) &= \mathbb{E}_{a\sim\pi_t}\left[f^t_l(\tau_t[l])\right] + \mathbb{E}_{a\sim\pi_t}\left[\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_l,a)}V^t_{l+1,\mathbb{P}}((\tau_t[l-1], s, a), s'))\right] \\ &= \mathbb{E}_{a\sim\pi_t}\left[f^t_l(\tau_t[l])\right] + \mathbb{E}_{a\sim\pi_t}\left[\mathbb{P}(\cdot \mid s_l, a)^\top V^t_{l+1,\mathbb{P}}((\tau_t[l-1], s_l, a), \cdot)\right] \end{aligned}$$

Now use τ_t to set $\tau_{l-1} := \tau_t[l-1]$ and define the following.

$$\Delta^t_l(s_l) := V^t_{l,\mathbb{P}_\star}(\tau_t[l-1], s_l) - V_{l,\widetilde{\mathbb{P}}_t}(\tau_t[l-1], s_l)$$

Note that

$$\Delta^t_1(s_1) = V(\widetilde{\mathbb{P}}_t, f^t, \pi_t) - V(\mathbb{P}_\star, f^t, \pi_t) \tag{1}$$

The computation above then gives us the following.

$$\begin{aligned} \Delta^t_l(s^t_l) &= \mathbb{E}_{a\sim\pi_t}\left[\widetilde{\mathbb{P}}_t(\cdot \mid s^t_l, a)^\top V^t_{l+1,\widetilde{\mathbb{P}}_t}((\tau_t[l-1], s^t_l, a), \cdot))\right] \\ &\quad - \mathbb{E}_{a\sim\pi_t}\left[\mathbb{P}_\star(\cdot \mid s^t_l, a)^\top V^t_{l+1,\mathbb{P}_\star}((\tau_t[l-1], s^t_l, a), \cdot))\right] \\ &= \widetilde{\mathbb{P}}_t(\cdot \mid s^t_l, a^t_l)^\top V^t_{l+1,\widetilde{\mathbb{P}}_t}(\tau_t[l], \cdot) - \mathbb{P}_\star(\cdot \mid s^t_l, a^t_l)^\top V^t_{l+1,\mathbb{P}_\star}(\tau_t[l], \cdot) + Y_{l,t} + Z_{l,t} \end{aligned}$$

where $Y_{l,t}$ and $Z_{l,t}$ are stochastic processes defined below.

$$Y_{l,t} := \mathbb{P}_\star(\cdot \mid s_l^t, a)^\top V^t_{l+1,\mathbb{P}_\star}(\tau_t[l], \cdot) - \mathbb{E}_{a\sim\pi_t}\left[\mathbb{P}_\star(\cdot \mid s_l^t, a)^\top V^t_{l+1,\mathbb{P}_\star}((\tau_t[l-1], s_l^t, a), \cdot)\right]$$
$$Z_{l,t} := \mathbb{E}_{a\sim\pi_t}\left[\widetilde{\mathbb{P}}_t(\cdot \mid s_l^t, a)^\top V^t_{l+1,\widetilde{\mathbb{P}}_t}((\tau_t[l-1], s_l^t, a), \cdot)\right] - \widetilde{\mathbb{P}}_t(\cdot \mid s_l^t, a)^\top V^t_{l+1,\widetilde{\mathbb{P}}_t}(\tau_t[l], \cdot))$$

Consider the filtration $\mathcal{G}_{l,t}$ induced by the data of $\{\tau_s\}_{s=1}^{t-1} \cup \tau_t[l-1] \cup \{s_l^t\}$. Since $a_l^t \sim \pi_t$ and $(\tau_t[l-1], s_l^t, a_l^t) = \tau_t[l]$, we get that $\mathbb{E}[Y_{l,t}|\mathcal{G}_{l,t}] = \mathbb{E}[Z_{l,t}|\mathcal{G}_{l,t}] = 0$. So, one can see that both processes are martingales over $\mathcal{G}_{l,t}$. Also note that $|Y_{l,t}|, |Z_{l,t}| \leqslant p$. We thus have that

$$\begin{aligned}
\Delta_l^t(s_l^t) &= \left[\widetilde{\mathbb{P}}_t(\cdot \mid s_l^t, a_l^t) - \mathbb{P}_\star(\cdot \mid s_l^t, a_l^t)\right] V^t_{l+1,\widetilde{\mathbb{P}}_t}(\tau_t[l], \cdot)) \\
&\quad + \mathbb{P}_\star(\cdot \mid s_l^t, a_l^t)^\top \left[V^t_{l+1,\widetilde{\mathbb{P}}_t}(\tau_t[l], \cdot)) - V^t_{l+1,\mathbb{P}_\star}(\tau_t[l], \cdot)\right] + Y_{l,t} + Z_{l,t} \\
&= \left[\widetilde{\mathbb{P}}_t(\cdot \mid s_l^t, a_l^t) - \mathbb{P}_\star(\cdot \mid s_l^t, a_l^t)\right] V^t_{l+1,\widetilde{\mathbb{P}}_t}(\tau_t[l], \cdot)) + \mathbb{P}_\star(\cdot \mid s_l^t, a_l^t)^\top \Delta_{l+1}^t(s') + Y_{l,t} + Z_{l,t} \\
&= \left[\widetilde{\mathbb{P}}_t(\cdot \mid s_l^t, a_l^t) - \mathbb{P}_\star(\cdot \mid s_l^t, a_l^t)\right] V^t_{l+1,\widetilde{\mathbb{P}}_t}(\tau_t[l], \cdot)) + \mathbb{E}_{s'\sim\mathbb{P}_\star(\cdot|s_l^t,a_l^t)}\left[\Delta_{l+1}^t(s')\right] + Y_{l,t} + Z_{l,t} \\
&= \left[\widetilde{\mathbb{P}}_t(\cdot \mid s_l^t, a_l^t) - \mathbb{P}_\star(\cdot \mid s_l^t, a_l^t)\right] V^t_{l+1,\widetilde{\mathbb{P}}_t}(\tau_t[l], \cdot)) + \Delta_{l+1}^t(s_{l+1}^t) + U_{l,t} + Y_{l,t} + Z_{l,t}
\end{aligned}$$

where

$$U_{l,t} := \mathbb{E}_{s'\sim\mathbb{P}_\star(\cdot|s_l^t,a_l^t)}\left[\Delta_{l+1}^t(s')\right] - \Delta_{l+1}^t(s_{l+1}^t)$$

Consider the filtration $\bar{\mathcal{G}}_{l,t}$ defined by the data of $\{\tau_s\}_{s=1}^{t-1} \cup \tau_t[l]$. Clearly, $U_{l,t}$ is a martingale over $\bar{\mathcal{G}}_{l,t}$. Also note that $|U_{l,t}| \leqslant p$ To conclude, we have that

$$\Delta_l^t(s_l^t) - \Delta_{l+1}^t(s_{l+1}^t) = \left[\widetilde{\mathbb{P}}_t(\cdot \mid s_l^t, a_l^t) - \mathbb{P}_\star(\cdot \mid s_l^t, a_l^t)\right] V^t_{l+1,\widetilde{\mathbb{P}}_t}(\tau_t[l], \cdot)) + U_{l,t} + Y_{l,t} + Z_{l,t}$$

Using a telescoping sum over l for a fixed t and equation 1, we get that for any t, the following holds.

$$\begin{aligned}
&V(\widetilde{\mathbb{P}}_t, f^t, \pi_t) - V(\mathbb{P}_\star, f^t, \pi_t) \\
&= \Delta_1^t(s_1) \\
&= \sum_{l=1}^H \left[\widetilde{\mathbb{P}}_t(\cdot \mid s_l^t, a_l^t) - \mathbb{P}_\star(\cdot \mid s_l^t, a_l^t)\right] V^t_{l+1,\widetilde{\mathbb{P}}_t}(\tau_t[l], \cdot)) + U_{l,t} + Y_{l,t} + Z_{l,t}
\end{aligned}$$

$$\begin{aligned}
&V(\widetilde{\mathbb{P}}_t, f^t, \pi_t) - V(\mathbb{P}_\star, f^t, \pi_t) \\
&\leqslant \sum_{l=1}^H Bp \left\|\widetilde{\mathbb{P}}_t(\cdot \mid s_l^t, a_l^t) - \mathbb{P}_\star(\cdot \mid s_l^t, a_l^t)\right\|_1 + U_{l,t} + Y_{l,t} + Z_{l,t} \\
&\leqslant \sum_{l=1}^H Bp \left\|\widetilde{\mathbb{P}}_t(\cdot \mid s_l^t, a_l^t) - \hat{\mathbb{P}}_t(\cdot \mid s_l^t, a_l^t)\right\|_1 + Bp \left\|\mathbb{P}_\star(\cdot \mid s_l^t, a_l^t) - \hat{\mathbb{P}}_t(\cdot \mid s_l^t, a_l^t)\right\|_1 + U_{l,t} + Y_{l,t} + Z_{l,t}
\end{aligned} \tag{2}$$

Until equation 2, all statements have held with probability 1 and did not use any facts about $\widetilde{\mathbb{P}}_t$. The last inequality also holds with probability 1 and uses the design of the confidence sets. Now, note the following well known concentration lemma. See, for example, [50].

Lemma 11. *For* $\zeta(n, \delta) = 8\sqrt{\frac{S\log(2)+\log(n(n+1)SA/\delta)}{2n}}$ *and*

$$\mathcal{C}_{\mathcal{P}}^t(\delta) = \left\{\mathbb{P} \middle| \|\mathbb{P}(\cdot \mid s, a) - \hat{\mathbb{P}}_t(\cdot \mid s, a)\|_1 \leqslant \zeta(N_t(s, a), \delta) \forall s, a\right\}$$

the true model $\mathbb{P}^\star \in \mathcal{C}_{\mathcal{P}}^t(\delta)$ *for all* $t \geqslant 1$ *with probability at least* $1 - \delta/32$.

Applying the lemma twice and applying a union bound imply that the following holds with probability $1-\delta/8$.

$$\begin{aligned}
&\sum_{t=1}^{T} V(\widetilde{\mathbb{P}}_t, f^t, \pi_t) - V(\mathbb{P}_\star, f^t, \pi_t) \\
&\overset{(i)}{\leqslant} \sum_{l=1}^{H} 2Bp\zeta(N_t(s_l^t, a_l^t), \delta) + U_{l,t} + Y_{l,t} + Z_{l,t} \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H} 2Bp\zeta(N_t(s_h^t, a_h^t), \delta) + \left[\sum_{t=1}^{T}\sum_{h\in\mathcal{H}_p} U_{h,t} + Y_{h,t} + Z_{h,t}\right] \\
&\overset{(ii)}{\leqslant} \sum_{t=1}^{T}\sum_{h=1}^{H} 2Bp\zeta(N_t(s_h^t, a_h^t), \delta) + \mathcal{O}\left(Bp\sqrt{pT\log(T/\delta)}\right)
\end{aligned}$$

Note that inequality (i) is subtle since we could have used more data than that from $\tau_i, i = 1 \to t$ to construct $\mathcal{C}_{\mathcal{P}}^t$. The inequality still holds since $\zeta(n, \delta)$ is decreasing in n. Also, inequality (ii) holds by the Azuma-Hoeffding inequality.

Now note that the whole argument above can be repeated with $\mathbb{P}_\star$ and $\widetilde{\mathbb{P}}_t$ switched, since the negative of a martingale is also a martingale. So, we have that with probability $1-\delta/4$

$$\left|\sum_{t=1}^{T} V(\widetilde{\mathbb{P}}_t, f^t, \pi_t) - V(\mathbb{P}_\star, f^t, \pi_t)\right| \leqslant \sum_{t=1}^{T}\sum_{h=1}^{H} 2Bp\zeta(N_t(s_h^t, a_h^t), \delta) + \mathcal{O}\left(Bp\sqrt{pT\log(T/\delta)}\right)$$

Finally, we need the following easy lemma, proved in [50].

Lemma 12. *Let $c_\delta := 8\sqrt{S\log(2) + \log(HTSA/\delta)}$. Then the following holds almost surely.*

$$\sum_{t=1}^{T}\sum_{h=1}^{H} 2Bp\zeta(N_t(s_h^t, a_h^t), \delta) \leqslant c_\delta Bp\sqrt{SAHT} + c_\delta BpHSA$$

This establishes our claim. □

D.2 Putting It All Together

Theorem 1 (POR-UCRL Regret)**.** *Under Assumption 1, the regret* $\mathrm{Regret}(T)$ *of POR-UCRL is bounded by the following with probability at least* $1-\delta$

$$\widetilde{\mathcal{O}}\left(\left(pS\sqrt{HA} + \sum\nolimits_{h\in\mathcal{H}_p}\sqrt{d_{E,h}d_{C,h}}\right)\sqrt{T}\right)$$

where $d_{E,h} = \dim_E\left(\mathcal{F}_h, \frac{B}{T}\right)$ *and* $d_{C,h} = \log(\mathcal{N}(\mathcal{F}_h, 1/T, \|\cdot\|_\infty))$.

Proof. We can now combine Lemmas 7and 11 to conclude that $\mathbb{M}_\star \in \mathcal{C}_{\mathcal{M}}(\mathcal{D}_t, \delta)$ for all t with probability $1-\delta/16$. We can now combine this observation with Lemmas 8 and 9 to observe that Assumption 3 is satisfied by POR-UCRL. By Theorem 4, the following holds with probability $1-\delta$.

$$\mathrm{Regret}(T) = \mathcal{O}\left(c_\delta Bp\sqrt{SAHT} + c_\delta BpHSA + \sum_{h\in\mathcal{H}_p} B\kappa_{2,h}d_{E,h} + \sum_{h\in\mathcal{H}_p}\kappa_{2,h}\sqrt{d_{E,h}\beta_{h,T}(\delta)T}\right)$$

where $c_\delta = 8\sqrt{S\log(2) + \log(HTSA/\delta)}$, $d_{E,h} = \dim_E\left(\mathcal{F}_h, \frac{B}{T}\right)$ is the Eluder dimension of $\mathcal{F}_h$ and $\beta_{h,T}(\delta) = \beta_{h,t}\left(\frac{\delta}{2t^2H} = \widetilde{O}(B^2\eta_h^2 d_{C,h})\right)$. This is because all the terms dependent on p get absorbed by the first term in our expression below.

We further refine it by ignoring terms independent of T and using the fact that $\beta_{h,T}(\delta) = \widetilde{\mathcal{O}}(d_{C,h})$ to get that

$$\text{Regret}(T) = \widetilde{\mathcal{O}}\left(pS\sqrt{AHT} + \sum_{h\in\mathcal{H}_p}\sqrt{d_{E,h}d_{C,h}T}\right)$$

□

Analogously, we can provide a sample complexity result for POR-UCRL.

Corollary 2 (POR-UCRL Sample complexity)**.** *Let $\varepsilon > 0, \delta \in [0,1]$. Ignoring polynomial terms independent of ε, we can bound the sample complexity $N(\varepsilon,\delta)$ of POR-UCRL as follows*

$$\widetilde{\mathcal{O}}\left(\frac{p^2HS^2A}{\varepsilon^2} + \frac{p^2d_Ed_C}{\varepsilon^2}\right)$$

where $d_E := \max_{h\in\mathcal{H}_p} d_{E,h}$, and $d_C := \max_{h\in\mathcal{H}_p} d_{C,h}$.

Proof. We invoke the regret-to-PAC conversion in Lemma 5 with confidence $\delta' = \delta/2$ and we plug the regret bound in Theorem 1 to write

$$\varepsilon = \widetilde{\mathcal{O}}\left(\left(BpS\sqrt{AH} + \sum_{h\in\mathcal{H}_p}\kappa_{2,h}\sqrt{d_{E,h}d_{C,h}} + Bp\sqrt{\log(1/\delta)}\right)\left(\frac{1}{\sqrt{T}}\right)\right)$$

from which we get the result by picking $N = T$ and the definition of d_E, d_C. □

Also note the following theorem and corresponding corollary.

Theorem 7 (POR-UCRL Regret if $\mathbb{P}_\star$ is Known)**.** *If we know the transition matrix $\mathbb{P}_\star$ in POR-UCRL, then our regret is given by the following with probability $1-\delta$, ignoring polynomial terms independent of T.*

$$\text{Regret}(T) = \widetilde{\mathcal{O}}\left(\left(Bp + \sum_{h\in\mathcal{H}_p}\sqrt{d_{E,h}d_{C,h}}\right)\sqrt{T}\right)$$

Proof. We can now use Lemmas 8 and the fact that $\mathcal{C}^t_{\mathcal{P}}(\delta)$ is always a singleton to observe that Assumption 3 is satisfied by this version of POR-UCRL as well. By Theorem 4, the following holds with probability $1-\delta$.

$$\text{Regret}(T) = \widetilde{\mathcal{O}}\left(Bp\sqrt{T} + \sum_{h\in\mathcal{H}_p}B\kappa_{2,h}d_{E,h} + \sum_{h\in\mathcal{H}_p}\kappa_{2,h}\sqrt{d_{E,h}\beta_{h,T}(\delta)T}\right)$$

We further refine it by ignoring terms independent of T and using the fact that $\beta_{h,T}(\delta) = \widetilde{\mathcal{O}}(d_{C,h})$ to get that

$$\text{Regret}(T) = \widetilde{\mathcal{O}}\left(\left(Bp + \sum_{h\in\mathcal{H}_p}\kappa_{2,h}\sqrt{d_{E,h}\beta_{h,T}(\delta)}\right)\sqrt{T}\right)$$

□

Corollary 3 (POR-UCRL sample complexity if $\mathbb{P}_\star$ is Known)**.** *Let $\varepsilon > 0, \delta \in [0,1]$. Ignoring polynomial terms independent of ε, we can bound the sample complexity $N(\varepsilon,\delta)$ of POR-UCRL when $\mathbb{P}_\star$ is known as follows*

$$\widetilde{\mathcal{O}}\left(\frac{p^2d_Ed_C}{\varepsilon^2}\right)$$

where $d_E := \max_{h\in\mathcal{H}_p} d_{E,h}$, and $d_{C,h} := \max_{h\in\mathcal{H}_p} d_{C,h}$.

Proof. The proof proceeds as in Corollary 2 by plugging Theorem 7 in Lemma 5. □

E Details and Proofs for Cardinal POR-UCBVI

We now describe how we instantiate POR-UCBVI from a generic optimistic algorithm using bonuses. Note that again, this is **crucially different** from naively summarizing the history to define a modified state space, since we are separating the use of history summarization for getting bonuses for f from using only the current state while getting bonuses for the Markovian transitions $\mathbb{P}$. Like with confidence sets, it is a priori unclear if we can use ideas from optimism to prove guarantees with a favorable (non-exponential) dependence on the complexity of transitions. In particular, we will note that showing that the bonuses are optimistic would naively need a union bound over the doubly exponential ($A^{(SA)^H}$) set of history-dependent policies, which is a non-trivial challenge to overcome.

Given a dataset of the first t trajectory samples $\{\tau_i\}_{i=1}^t$ and an index $h \in [H]$, we consider the following:

Estimates for POR-UCBVI:

$$\widehat{f}_h^{t+1} = \arg\min_{f_h \in \mathcal{F}_h} \sum_{i=1}^{t} \left(\sigma(f_h(\tau_i[h])) - o_h^i\right)^2$$

We also use the MLE estimate for $\mathbb{P}$ after t episodes to define $\hat{\mathbb{P}}^t(\cdot \mid s,a) := \frac{N_t(s,a,s')}{N_t(s,a)}$. Now for $\zeta(n,\delta) = 2\sqrt{\frac{S\log(2)+\log(n(n+1)SA/\delta)}{2n}}$, define $\mathcal{C}_{\mathbb{P}_t}(\delta)$ as below:

$$\left\{\mathbb{P}\middle|\|\mathbb{P}(\cdot \mid s,a) - \hat{\mathbb{P}}_t(\cdot \mid s,a)\|_1 \leqslant \zeta(N_t(s,a),\delta)\forall s,a\right\}$$

Recall the definition of our bonus below.

Bonuses for POR-UCBVI. Recall that simple least squares guarantees imply the lemma below.

Lemma 7 (Concentration for $\sigma \circ f_h$)**.** *Define*

$$\mathrm{MSE}_{h,t}(f_h, f_h') := \sum_{i=1}^{t} \left(\sigma_h(f_h(\tau_i[h])) - \sigma_h(f_h'(\tau_i[h]))\right)^2$$

Also define $\bar{\beta}_{h,t}(\delta) = \eta_h^2 \log\left(\frac{N\left(\mathcal{F}_h, \frac{B}{T}, \|\cdot\|_\infty\right)}{\delta}\right) + \alpha_{h,t}$ *with* $\alpha_{h,t} := \frac{tB + t\eta_h \log\left(\frac{t}{\delta}\right)}{T}$*. Then* $f_h^\star$ *simultaneously satisfies* $\mathrm{MSE}_{h,t}(f_h^\star, \widehat{f}_h^{(t+1)}) \leqslant \bar{\beta}_{h,t}\left(\frac{\delta}{2t^2H}\right)$ *for all* h, t *with probability* $1 - \delta/32$.

We use the data from trajectories $\{\tau_i\}_{i=1}^t$ to build the confidence sets $\mathcal{C}_{\mathcal{F}}^{t+1}(\delta) = \prod_h \mathcal{C}_h^{t+1}(\delta)$ with $\mathcal{C}_h^{t+1}(\delta)$ defined below, where $\beta_{h,t}(\delta) := \bar{\beta}_{h,t}\left(\frac{\delta}{2t^2H}\right)$.

$$\mathcal{C}_h^{t+1}(\delta) := \left\{f_h \in \mathcal{F}_h \middle| \mathrm{MSE}_{h,t}(f_h^\star, \widehat{f}_h^{(t+1)}) \leqslant \beta_{h,t}(\delta)\right\}$$

We first define a trajectory dependent bonus term below, with $\bar{\delta} := \frac{\delta}{HS^HA^H}$

$$\gamma_{h,t}(\tau[h],\delta) = \max_{f_h, f_h' \in \mathcal{C}_h^t(\bar{\delta})} f_h(\tau[h]) - f_h'(\tau[h])$$

Note that according to the definition of β, this does not create any exponential dependence in the confidence intervals used to define $\mathcal{C}_h^{t+1}$.

$$\begin{aligned}\beta_{h,t}\left(\frac{\delta}{16S^HA^H}\right) &\leqslant 64\left(\log(N(\mathcal{F}_h, \alpha, \|\cdot\|_\infty)) + B + \eta_h\log(1/\delta) + \eta_h^2 H\log(THSA/\delta)\right)\\ &= O(d_{C,h} + H)\end{aligned}$$

It follows by a union bound over all trajectory segments and all timesteps t that with probability at least $1 - \delta/16$ and for any trajectory τ and $t \geqslant 1, h \in \mathcal{H}_p$,

$$|f_h^\star(\tau[h]) - \hat{f}_h^t(\tau[h])| \leqslant \gamma_{h,t}(\tau[h],\delta) \tag{3}$$

Remark 4. In the case of many popular function classes $\mathcal{F}$, like the linear class $\mathcal{F}_H = \{\tau \mapsto \phi(\tau)^\top \mathbf{w} \mid \|\mathbf{w}\| \leqslant W\}$, we can compute $\gamma_{h,t}(\tau[h], \delta)$ quite easily. In this case $\gamma_{H,t}$ is given by

$$\sup_{\mathbf{w},\mathbf{w}' \in W_t} \phi(\tau)^\top (\mathbf{w} - \mathbf{w}') = \|\phi(\tau)\|_{V_t} \sup_{\mathbf{w},\mathbf{w}' \in W_t} \|w - w'\|_{V_t^{-1}}$$

for a suitable quadratic form V_t.

$\gamma_{h,t}(\tau[h], \delta)$ induces a trajectory-dependent bonus, given by

$$b^t_{\mathcal{F}}(\tau, \delta) := \sum_{h \in \mathcal{H}_p} \gamma_{h,t}(\tau[h], \delta)$$

This in turn induces a policy-level bonus (which depends on the transition kernel), given by:

$$b^t_{\mathcal{F}}(\mathbb{P}, \pi, \delta) := \mathbb{E}_{\tau \sim \mathbb{P}^\pi_\star}\left[b^t_{\mathcal{F}}(\tau, \delta)\right] = \mathbb{E}_{\tau \sim \mathbb{P}^\pi}\left[\sum_{h \in \mathcal{H}_p} \gamma_{h,t}(\tau[h], \delta)\right]$$

Let us define a term $\xi^t(s, a, \delta)$ that will be used to define the probability bonus.

$$\xi^t(s, a, \delta) := \min\left(2, 4\sqrt{\frac{H\log(6HSA) + S\log(8t^2H^2) + \log(32t^2 N_t(s,a)/\delta)}{2N_t(s,a)}}\right)$$

This induces a trajectory-dependent bonus, given by

$$b^t_{\mathcal{P}}(\tau, \delta) := \sum_{h=1}^{H-1} \xi^t(s_h, a_h, \delta)$$

This induces a policy-level bonus (which depends on the transition kernel), given by:

$$b^t_{\mathcal{P}}(\mathbb{P}, \pi, \delta) := \min\left(4, \mathbb{E}_{\tau \sim \mathbb{P}^\pi_\star}\left[b^t_{\mathcal{F}}(\tau, \delta)\right]\right) = \min\left(4, \mathbb{E}_{\tau \sim \mathbb{P}^\pi}\left[\sum_{h=1}^{H-1} \xi^t(s_h, a_h, \delta)\right]\right)$$

Estimates and Bonuses in case $\mathbb{P}_\star$ is known. If $\mathbb{P}_\star$ is instead known, keep $\hat{f}^t$ and $b^t_{\mathcal{F}}(\mathbb{P}, \pi, \delta)$ the same as above, but set $\hat{\mathbb{P}}_t := \mathbb{P}_\star$ and $b^t_{\mathcal{P}}(\mathbb{P}, \pi, \delta) := 0$ for all t.

For completeness we state POR-UCBVI here, which is an instantiation of Algorithm 3, the generic optimistic algorithm using bonuses.

Algorithm 6 POR-UCBVI

1: **Input** Known family of reward functions $\{\mathcal{R}_h\}_{h=1}^H$, methods $\texttt{Est}(\mathcal{D})$ to estimate $\hat{\mathbb{P}}_t$ and $\hat{f}^t$ from dataset $\mathcal{D}$, confidence level δ
2: **Initialize** $\mathcal{D}_1 \leftarrow \{\}$, initialize $\hat{f}^{\mathcal{D}_1}, \hat{\mathbb{P}}_{\mathcal{D}_1}$ arbitrarily.
3: **for** $t = 1, ..., T$ **do**
4: **Compute** optimistic history dependent policy,

$$\pi_t = \arg\max_\pi V(\hat{\mathbb{P}}_t, \hat{f}^t, \pi) + b^t_{\mathcal{F}}(\hat{\mathbb{P}}_t, \pi, \delta) + z(Bp)(b^t_{\mathcal{P}}(\hat{\mathbb{P}}_t, \pi, \delta))$$

5: **Observe** trajectory $\tau_t = \{(s^t_h, a^t_h)\}_{h=1}^H$ and feedback $\{o_h\}_{h \in \mathcal{H}_p}$.
6: **Compute** new estimates $\hat{f}^{t+1}, \hat{\mathbb{P}}_{t+1} \leftarrow \texttt{Est}(\mathcal{D}_{t+1})$ and compute new bonus functions $b^{t+1}_{\mathcal{F}}(\hat{f}^{t+1}, \cdot, \delta), b^{t+1}_{\mathcal{P}}(\hat{\mathbb{P}}_{t+1}, \cdot, \delta)$.
7: **end for**

We will show the following regret bound.

Theorem 8 (POR-UCBVI Regret)**.** *Under Assumption 1, POR-UCBVI satisfies Assumption 4 and its regret* $\mathrm{Regret}(T)$ *is bounded by the following with probability at least* $1-\delta$*, ignoring polynomial terms independent of* T*.*

$$\widetilde{\mathcal{O}}\left(\left(pC(H,S,A)+\sum_{h\in\mathcal{H}_p}\sqrt{d_{E,h}(d_{C,h}+H)}\right)\sqrt{T}\right)$$

where $C(H,S,A) := H\sqrt{SA}+S\sqrt{HA}$

E.1 Showing that Assumption 4 is Satisfied

E.1.1 Bounding effect of error in $\mathcal{F}$

Lemma 13 (Bounding $\hat{f}^t$ Value Error)**.** *Given any* $\mathbb{P}$*, with* $\hat{f}^t$ *computed using data from* $\{\tau_i\}_{i=1}^t \sim \mathbb{P}_\star^{\pi_i}$ *for any sequence of policies* π_i *using least squares, the following holds with probability* $1-\delta/16$ *uniformly over all* π*.*

$$|V(\mathbb{P},\hat{f}^t,\pi)-V(\mathbb{P},f^\star,\pi)| \leqslant b^t_{\mathcal{F}}(\mathbb{P},\pi,\delta)$$

Proof. Recall that with probability at least $1-\delta/16$, the following holds for any trajectory τ and any $t\geqslant 1, h\in\mathcal{H}_p$.

$$|f^\star_h(\tau[h])-\hat{f}^t_h(\tau[h])| \leqslant \gamma_{h,t}(\tau[h],\delta) \tag{4}$$

Now note the following inequalities, where (i) holds with probability $1-\delta/16$ uniformly over all policies due to inequality 4 above.

$$\begin{aligned}
V(\mathbb{P},\hat{f}^t,\pi)-V(\mathbb{P},f^\star,\pi) &= \mathbb{E}_{\tau\sim\mathbb{P}^\pi}\left[\sum_{h\in\mathcal{H}_p}\hat{f}^t_h(\tau[h])-f^\star_h(\tau[h])\right]\\
&\overset{(i)}{\leqslant} \mathbb{E}_{\tau\sim\mathbb{P}^\pi}\left[\sum_{h\in\mathcal{H}_p}\gamma_{h,t}(\tau[h],\delta)\right]\\
&= b^t_{\mathcal{F}}(\mathbb{P},\pi,\delta)
\end{aligned}$$

□

Lemma 14 (Bounding Sum of $\mathcal{F}$ Bonuses)**.** *The following holds with probability* 1*.*

$$\sum_{t=1}^T b^t_{\mathcal{F}}(\mathbb{P}_\star,\pi_t,\delta) = \widetilde{\mathcal{O}}\left(\sum_{h\in\mathcal{H}_p}Bd_{E,h}+\sum_{h\in\mathcal{H}_p}\sqrt{d_{E,h}\beta_{h,T}(\bar{\delta})T}+Bp\sqrt{T\log(T/\delta)}\right)$$

Proof. First note the following inequality, which hold with probability $1-\delta/16$ by the Azuma-Hoeffding inequality.

$$\begin{aligned}
\sum_{t=1}^T b^t_{\mathcal{F}}(\mathbb{P}_\star,\pi_t,\delta) &= \sum_{t=1}^T \mathbb{E}_{\tau\sim\mathbb{P}_\star^{\pi_t}}\left[\sum_{h\in\mathcal{H}_p}\gamma_{t,h}(\tau[h],\delta)\right]\\
&\leqslant \sum_{t=1}^T\sum_{h\in\mathcal{H}_p}\gamma_{t,h}(\tau[h],\delta)+\mathcal{O}\left(Bp\sqrt{T\log(T/\delta)}\right)
\end{aligned}$$

Now apply Lemma 3 in [10] for each h separately, with the function class in the lemma set to $\{\sigma\circ f_h | f_h\in\mathcal{F}_h\}$, $P=1$, $\mathbf{x}_{t,p}=\mathbf{x}_{t,1}=\tau_t[h]$ and misspecification $\varepsilon=0$ (decoupled from the

Eluder dimension's ε). We also note that o_h^t are η-subgaussian samples with mean $\sigma(f_h(\tau[h]))$. We obtain

$$\sum_{t=1}^{T} \max_{f_h, f_h' \in \mathcal{C}_h^t(\delta)} \sigma(f_h(\tau[h])) - \sigma(f_h'(\tau[h])) \leqslant \mathcal{O}\left(B d_{E,h} + \sqrt{d_{E,h}\beta_{h,T}(\bar{\delta})T}\right)$$

where $d_{E,h} = \dim_E\left(\mathcal{F}_h, \frac{B}{T}\right)$ is the Eluder dimension of $\mathcal{F}_h$ and $\beta_{h,T}(\bar{\delta}) = \bar{\beta}\left(\frac{\bar{\delta}}{2t^2 H}\right)$. Since the Lipschitz constant of σ^{-1} is κ_2, we have that the following holds with probability 1.

$$\sum_{t=1}^{T} \gamma_{t,h}(\tau[h], \delta) \leqslant \mathcal{O}\left(B\kappa_2 d_{E,h} + \kappa_2\sqrt{d_{E,h}\beta_{h,T}(\bar{\delta})T}\right)$$

This implies that the following holds with probability $1 - \delta/16$.

$$\begin{aligned}\sum_{t=1}^{T} b_{\mathcal{F}}^t(\mathbb{P}_\star, \pi_t, \delta) &\leqslant \sum_{t=1}^{T}\sum_{h\in\mathcal{H}_p} \gamma_{t,h}(\tau[h], \delta) + \mathcal{O}\left(Bp\sqrt{T\log(T/\delta)}\right) \\ &= \widetilde{\mathcal{O}}\left(\sum_{h\in\mathcal{H}_p} B\kappa_2 d_{E,h} + \sum_{h\in\mathcal{H}_p}\kappa_2\sqrt{d_{E,h}\beta_{h,T}(\bar{\delta})T} + Bp\sqrt{T\log(T/\delta)}\right)\end{aligned}$$

□

E.1.2 Bounding effect of error in $\mathcal{P}$

We now restate Lemma B.2 of [11] in our notation.

Lemma 15 (Change of Measure Inequality)**.** *For any function μ of trajectories bounded by D, if $\hat{\mathbb{P}}_t$ is computed from data that includes trajectories $\{\tau_i \sim \mathbb{P}_\star^{\pi_i}\}_{i=1}^t$ for any sequence of policies π_i, then the following holds uniformly over all policies π with probability $1 - \delta/16$.*

$$\mathbb{E}_{\tau\sim\mathbb{P}_\star}[\mu(\tau)] - \mathbb{E}_{\tau\sim\hat{\mathbb{P}}_t}[\mu(\tau)] \leqslant 2D\sqrt{\log(D)}\, b_{\mathcal{P}}^t(\hat{\mathbb{P}}_t, \pi, \delta)$$

The same statement holds if we switch the roles of $\mathbb{P}$ and $\hat{\mathbb{P}}_t$ on both sides.

Proof. For the order of $\mathbb{P}$ and $\hat{\mathbb{P}}_t$ in the statement, the following follows from Lemma B.2 of [11] with $\eta = D$ and $\varepsilon = \frac{1}{t^2}$. We pull the additive $\log(D)$ in the square root outside to fit our assumption's phrasing.

$$\mathbb{E}_{\tau\sim\mathbb{P}_\star}[\mu(\tau)] - \mathbb{E}_{\tau\sim\hat{\mathbb{P}}_t}[\mu(\tau)] \leqslant D\sqrt{\log(D)}\, b_{\mathcal{P}}^t(\hat{\mathbb{P}}_t, \pi, \delta) + \frac{1}{t^2} \leqslant 2D\sqrt{\log(D)}\, b_{\mathcal{P}}^t(\hat{\mathbb{P}}_t, \pi, \delta)$$

The only subtlety is that more data than that from $\{\tau_i\}_{i=1}^t$ could have been used to compute $\hat{\mathbb{P}}_t$. The proof still follows since $c_{\mathcal{P}}(\hat{\mathbb{P}}_t, \pi, D)$ is decreasing in the counts $N_t(s, a)$.

Finally, if we switch $\mathbb{P}$ and $\hat{\mathbb{P}}_t$ on both sides, we can follow the proof of Lemma B.2 verbatim with $\mathbb{P}$ and $\hat{\mathbb{P}}_t$ switched everywhere, except for the martingale argument. There, instead of switching the two transition kernels, we negate the martingale to get our desired result. This exception is because we still need the expectation to be over the true transition kernel $\mathbb{P}_\star$ for the stochastic process defined to be a martingale. □

Lemma 16 (Bounding $\hat{\mathcal{P}}_t$ Value Error)**.** *Consider any sequence of functions f^t that induce value functions $V(\mathbb{P}, f^t, \pi)$. For any sequence of policies π_t, if the estimates $\hat{\mathbb{P}}_t$, bonuses $b_{\mathcal{P}}$ and costs $c_{\mathcal{P}}$ are generated using data including that of $\tau_i \sim \mathbb{P}_\star^{\pi_i}, i = 1 \to t$, then the following holds uniformly over t and over all policies with probability $1 - \delta/16$.*

$$V(\hat{\mathbb{P}}_t, f^t, \pi_t) - V(\mathbb{P}_\star, f^t, \pi_t) = Bp\sqrt{\log(Bp)}(b_{\mathcal{P}}^t(\mathbb{P}_\star, \pi, \delta))$$

The statement also holds if we switch $\hat{\mathbb{P}}_t$ and $\mathbb{P}_\star$.

Proof. Note the following two inequalities that immediately follow from Lemma 15

$$V(\mathbb{P}_\star, f^t, \pi_t) - V(\hat{\mathbb{P}}_t, f^t, \pi_t) \leqslant Bp\sqrt{\log(Bp)}(b^t_{\mathcal{P}}(\hat{\mathbb{P}}_t, \pi, \delta))$$
$$V(\hat{\mathbb{P}}_t, f^t, \pi_t) - V(\mathbb{P}_\star, f^t, \pi_t) \leqslant Bp\sqrt{\log(Bp)}(b^t_{\mathcal{P}}(\mathbb{P}_\star, \pi, \delta))$$

Our result follows immediately. □

Lemma 17 (Bounding Sum of $\mathcal{P}$ Bonuses)**.** *The following holds with probability* $1 - \delta/16$ *whenever the data used to compute* $b^t_{\mathcal{P}}(\mathbb{P}, \pi, \delta)$ *includes the data of trajectories* $\tau_i, t = 1 \to t$*.*

$$\sum_{t=1}^{T} b^t_{\mathcal{P}}(\mathbb{P}_\star, \pi_t, \delta) = \widetilde{\mathcal{O}}\left(SA\bar{c}_\delta + \bar{c}_\delta\sqrt{HSAT}\right)$$

where

$$\bar{c}_\delta := 4\sqrt{\frac{H\log(6HSA) + S\log(8t^2H^2) + \log(32t^2N_T(s,a)/\delta)}{2}}$$

This means that for any s, a*,* $\xi^t(s, a, \delta) = 2$ *until* $N_t(s, a) \geqslant \frac{\bar{c}_\delta}{2}$

Proof. First note that by the definition of the bonus and the Azuma-Hoeffding inequality, we have the following.

$$\begin{aligned}
\sum_{t=1}^{T} b^t_{\mathcal{P}}(\mathbb{P}_\star, \pi_t, \delta) &\leqslant \sum_{t=1}^{T} \mathbb{E}_{\tau\sim\mathbb{P}^\pi} b^t_{\mathcal{P}}(\tau, \delta) \\
&\leqslant \sum_{t=1}^{T} b^t_{\mathcal{P}}(\tau_t, \delta) + \mathcal{O}(4\sqrt{T\log(T/\delta)}) \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H-1} \xi^t(s^t_h, a^t_h, \delta) + \mathcal{O}(4\sqrt{T\log(T/\delta)})
\end{aligned}$$

Now note that the first inequality holds even if more data beyond that of $\{\tau_i\}_{i=1}^t$ is used to compute $\xi^t(s, a, \delta)$, since $\xi^t(s, a, \delta)$ is decreasing in $N_t(s, a)$.

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{h=1}^{H-1} \xi^t(s^t_h, a^t_h, \delta) &\leqslant SA\bar{c}_\delta + \sum_{t=1}^{T}\sum_{h=1}^{H-1} \frac{\bar{c}_\delta}{\sqrt{N_t(s^{(t)}_h, a^{(t)}_h)}} \\
&\leqslant SA\bar{c}_\delta + \bar{c}_\delta \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{l=1}^{N_T(s,a)} \frac{1}{\sqrt{l}} \\
&\leqslant SA\bar{c}_\delta + 2\bar{c}_\delta \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sqrt{N_T(s,a)} \\
&\leqslant SA\bar{c}_\delta + 2\bar{c}_\delta \sqrt{SA \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} N_T(s,a)} \\
&= \mathcal{O}\left(SA\bar{c}_\delta + \bar{c}_\delta\sqrt{SATH}\right)
\end{aligned}$$

This concludes our proof, since $1 = \widetilde{\mathcal{O}}(\sqrt{HSA})$ □

E.1.3 Putting everything together

Theorem 8 (POR-UCBVI Regret)**.** *Under Assumption 1, POR-UCBVI satisfies Assumption 4 and its regret* $\mathrm{Regret}(T)$ *is bounded by the following with probability at least* $1 - \delta$*, ignoring polynomial terms independent of* T*.*

$$\widetilde{\mathcal{O}}\left(\left(pC(H,S,A) + \sum_{h\in\mathcal{H}_p}\sqrt{d_{E,h}(d_{C,h} + H)}\right)\sqrt{T}\right)$$

where $C(H, S, A) := H\sqrt{SA} + S\sqrt{HA}$

Proof. Note that by Lemmas 13, 14, 15, 16, 17, Assumption 4 is satisfied by POR-UCBVI. Using Theorem 5 and Lemmas 14 and 17, we have the following.

$$\mathrm{Regret}(T) = \mathcal{O}\Big(\sum_{h\in\mathcal{H}_p} B\kappa_{2,h} d_{E,h} + BpHSA\bar{c}_\delta + \sum_{h\in\mathcal{H}_p} \kappa_{2,h}\sqrt{d_{E,h}\beta_{h,T}(\bar{\delta})T} + Bp\sqrt{T\log(T/\delta)} + \bar{c}_\delta BpH\sqrt{SAT}\Big)$$

We further refine it grouping terms and ignoring terms independent of T, and also noting that $\bar{c}_\delta = \widetilde{\mathcal{O}}(\sqrt{H}+\sqrt{S})$ as well as $\beta_{h,T}(\bar{\delta}) = \mathcal{O}(d_{C,h}+H)$

$$\begin{aligned}\mathrm{Regret}(T) &= \widetilde{\mathcal{O}}\left(\left(\sum_{h\in\mathcal{H}_p} \kappa_2\sqrt{d_{E,h}(d_{C,h}+H)} + Bp(H\sqrt{SA}+S\sqrt{HA})\right)\sqrt{T}\right)\\ &= \mathcal{O}\left(\left(p(H\sqrt{SA}+S\sqrt{HA}) + \sum_{h\in\mathcal{H}_p}\sqrt{d_{E,h}(d_{C,h}+H)}\right)\sqrt{T}\right)\end{aligned}$$

□

From the latter we derive a sample complexity result as follows.

Corollary 4 (POR-UCBVI Sample complexity)**.** *Let $\varepsilon > 0, \delta \in [0,1]$. Ignoring polynomial terms independent of ε, we can bound the sample complexity $N(\varepsilon,\delta)$ of POR-UCBVI as follows*

$$\widetilde{\mathcal{O}}\left(\frac{p^2 HSA\max(H,S)}{\varepsilon^2} + \frac{p^2 d_E \max(d_C,H)\log(1/\delta)}{\varepsilon^2}\right)$$

where $d_E := \max_{h\in\mathcal{H}_p} d_{E,h}$, and $d_C := \max_{h\in\mathcal{H}_p} d_{C,h}$.

Proof. We invoke the regret-to-PAC conversion in Lemma 5 with confidence $\delta' = \delta/2$ and we plug the regret bound in Theorem 8 to write

$$\varepsilon = \widetilde{\mathcal{O}}\left(\left(Bp(H\sqrt{SA}+S\sqrt{HA}) + \sum_{h\in\mathcal{H}_p}\sqrt{d_{E,h}\beta_{h,T}(\bar{\delta})} + Bp\sqrt{\log(1/\delta)}\right)\left(\frac{1}{\sqrt{T}}\right)\right)$$

from which we get the result by noting $N = pT$ and the definition of d_E, d_C. □

We also have the following theorem and corollary, in the same vein as Theorem 7.

Theorem 9 (Regret for POR-UCBVI if $\mathbb{P}_\star$ is Known)**.** *When $\mathbb{P}_\star$ is known, POR-UCBVI that sets $\hat{\mathbb{P}}_t := \mathbb{P}_\star$ and $b^t_{\mathcal{P}}(\mathbb{P},\pi\delta) := 0$ for all $t \geqslant 1$ still satisfies Assumption 4 and its regret $\mathrm{Regret}(T)$ is bounded by the following with probability at least $1-\delta$, ignoring terms independent of T.*

$$\widetilde{\mathcal{O}}\left(\left(\sum_{h\in\mathcal{H}_p}\sqrt{d_{E,h}(d_{C,h}+H)}\right)\sqrt{T}\right)$$

where $d_{E,h} = \dim_E\left(\mathcal{F}_h, \frac{B}{T}\right)$.

Proof. Note that by Lemmas 13 and 14, Assumption 4 is satisfied by POR-UCBVI. Using Lemma 14, we have the following.

$$\mathrm{Regret}(T) = \mathcal{O}\left(\sum_{h\in cH_p} B\kappa_2 d_{E,h} + \sum_{h\in\mathcal{H}_p}\kappa_2\sqrt{d_{E,h}\beta_{h,T}(\delta)T} + Bp\sqrt{T\log(T/\delta)}\right)$$

We further refine it grouping terms and ignoring terms independent of T, and also noting that $\bar{c}_\delta = \widetilde{\mathcal{O}}(\sqrt{H} + \sqrt{S})$

$$\begin{aligned}
\mathrm{Regret}(T) &= \widetilde{\mathcal{O}}\left(\left(\sum_{h\in\mathcal{H}_p} \kappa_2\sqrt{d_{E,h}\beta_{h,T}(\delta)} + Bp\right)\sqrt{T}\right) \\
&= \mathcal{O}\left(\left(Bp + \sum_{h\in\mathcal{H}_p} \sqrt{d_{E,h}\beta_{h,T}(\delta)}\right)\sqrt{T}\right) \\
&= \widetilde{\mathcal{O}}\left(\left(\sum_{h\in\mathcal{H}_p} \sqrt{d_{E,h}(d_{C,h} + H)}\right)\sqrt{T}\right)
\end{aligned}$$

□

Corollary 5 (POR-UCBVI Sample complexity if $\mathbb{P}_\star$ is Known)**.** *Let $\varepsilon > 0, \delta \in [0,1]$. Ignoring polynomial terms independent of ε, we can bound the sample complexity $N(\varepsilon,\delta)$ of POR-UCBVI when $\mathbb{P}_\star$ is known as follows*

$$\widetilde{\mathcal{O}}\left(\frac{p^2 H d_E \max(d_C, H)}{\varepsilon^2}\right)$$

where $d_E := \max_{h\in\mathcal{H}_p} d_{E,h}$, $\beta := \max_{h\in\mathcal{H}_p} \beta_{T,h}(\delta)$, and $d_C := \max_{h\in\mathcal{H}_p} d_{C,h}$.

Proof. The proof proceeds as in Corollary 4 by plugging Theorem 9 in Lemma 5. □

F Details and Proofs for PORRL with GOLF

For completeness and establishing notation, we recall GOLF here.

Algorithm 7 GOLF

Input Known class of Bellman consistent Q-functions $\mathcal{Q}$, confidence level δ.
Initialize dataset $\mathcal{D}_1 \leftarrow \{\}$ and $\mathcal{C}_{\mathcal{Q}}(\mathcal{D}_1, \delta) \leftarrow \mathcal{Q}$.
for $t = 1, ..., T$ **do**
  $\tau[0] \leftarrow ()$
  for $h = 1, \ldots H$ **do**
    Compute $a_h^t, Q_h^t \leftarrow \arg\max_{a, Q \in \mathcal{C}_{\mathcal{Q}}(\mathcal{D}_t, \delta)} Q(\tau[h], a)$
    Play a_h^t and observe feedback o_h^t
  end for
  Update $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup \{\tau, (o_1^t, \ldots o_H^t\}$
  Compute

$$\mathcal{C}_{\mathcal{Q}}(\mathcal{D}_{t+1}, \delta) \leftarrow \left\{ \mathcal{L}_{\mathcal{D}_t}(Q_h, Q_{h+1}) \leqslant \inf_{g \in \mathcal{G}_h} \mathcal{L}_{\mathcal{D}_t}(g, Q_{h+1}) + \beta \right\}$$

end for

Theorem 2 (Modified GOLF Regret)**.** *Let Assumption 1 hold, and let* $d_{\mathrm{HABE}} = \dim_{\mathrm{HABE}}(\mathcal{Q}, \alpha, \min(\alpha, \sqrt{1/T}))$. *Choose hyperparameter* $\beta = c\log(HT\mathcal{N}(\mathcal{Q} \cup \mathcal{G}, 1/T, \|\cdot\|_\infty))$ *for some universal constant* c *and the auxiliary function class* $\mathcal{G}$ *used in GOLF, and define* $d_{C,\mathcal{Q}} := \log(\mathcal{N}(\mathcal{Q} \cup \mathcal{G}, 1/T, \|\cdot\|_\infty))$. *Then, GOLF satisfies* $\mathrm{Regret}(T) = \mathcal{O}\left(pH\sqrt{d_{\mathrm{HABE}} d_{C,\mathcal{Q}} T}\right)$.

Proof. The meat of the theorem is in proving Lemma 18. We $\beta = c\log(HT\mathcal{N}(\mathcal{Q} \cup \mathcal{G}, 1/T, \|\cdot\|_\infty))$ for some suitably large universal constant c, and use Theorem 6 and Lemma 18 to get that

$$\mathrm{Regret}(T) = \sum_{h=1}^{H} \left(T\omega + Bp\left(\frac{B^2p^2\beta}{\alpha^2} + 1\right)\left(\sum_{l=1}^{h-1} d_l(\alpha)\right) + \min(d_h(\omega), T)Bp + 2Bp\sqrt{\beta d_h(\omega) T} \right)$$

where $d_h(\varepsilon) := \dim_{DE}(\Phi_h, \mathcal{D}_h, \varepsilon)$. Now set $\omega = \frac{Bp}{T}$ and use the fact that $d_h(\varepsilon)$ increases with decreasing ε to get that

$$\begin{aligned} \mathrm{Regret}(T) &= \widetilde{\mathcal{O}}\left(pH\sqrt{d_{\mathrm{HABE}}\beta T}\right) \\ &= \widetilde{\mathcal{O}}\left(pH\sqrt{d_{\mathrm{HABE}} d_{C,\mathcal{Q}} T}\right) \end{aligned}$$

since $d_{\mathrm{HABE}} := \dim_{\mathrm{HABE}}(\mathcal{Q}, \min(\alpha, Bp/T)) := \max_h d_h(\min(\alpha, Bp/T))$. □

Corollary 6 (GOLF Sample complexity)**.** *Let* $\varepsilon > 0, \delta \in [0, 1]$. *Ignoring polynomial terms independent of* ε, *we can bound the sample complexity* $N(\varepsilon, \delta)$ *of GOLF as follows*

$$\widetilde{\mathcal{O}}\left(\frac{p^2H^2 d_{\mathrm{HABE}} d_{c,\mathcal{Q}}}{\varepsilon^2}\right).$$

Proof. Again, we use Lemma 5 and a quick computation shows our result. □

F.1 Comparing $\dim_{\mathrm{HABE}}$ and $\dim_{\mathrm{BE}}$

It is easy to see that since the function class Φ_h is a subset of the class Ψ_h of all Bellman errors, $\dim_{\mathrm{HABE}} \leqslant \max_h \dim_{DE}(\Psi_h, \mathcal{D}_{h,\mathcal{Q}}, \varepsilon)$. Recall that the Bellman eluder dimension is a minimum over the RHS and another term that uses Dirac-δ distributions, but typically, the RHS is smaller. So, in many cases, $\dim_{\mathrm{HABE}} \leqslant \dim_{\mathrm{BE}}$. However, we don't have a universal inequality in either direction.

F.2 Computing dimensions for the combination lock

Proposition 1 (Dimensions for the Combination Lock)**.** *Consider the combination lock problem with model class $\mathcal{M} = \mathcal{P} \times \mathcal{F}$ and induced Q-function class $\mathcal{Q}$.*

- *Under dense intermediate feedback with $\mathcal{H}_p = [H]$, the $\dim_{\mathrm{HABE}}(\mathcal{Q}, \alpha)A$ for all $\alpha < q$, while its BE dimension is at least $A^H - 2$. The eluder dimension for reward functions $\dim_E(\mathcal{F}_h, \frac{B}{T})$ is at least A^h for any $h \leqslant H$.*
- *For sparse intermediate feedback with $\mathcal{H}_p = \{H\}$ and any $\alpha > 0$, the α-HABE dimension, the BE dimension and the eluder dimension of $\mathcal{F}_H$ are all at least $A^H - 2$.*

Proof. We separately resolve the cases of sparse and dense intermediate feedback.

F.2.1 Dense intermediate feedback, $\mathcal{H}_p = [H]$

Notice that we get a reward $Ber(q)$ at every step as long as we are on the correct sequence of actions $a_1^\star, \dots a_H^\star$, and as soon as we take a wrong action, we always get a reward of 0 subsequently. It is then easy to see that the induced function classes $\mathcal{Q}$ then are given by $\mathcal{Q} = \{(Q_1, \dots Q_H) \mid \exists a_1, \dots a_H \in \mathcal{A} \text{ s.t. } Q_h = (H - h + 1)q\mathbb{1}_{a_1, \dots a_h}\}$.

α-HABE dimension: It suffices to show the upper bound using $\mathcal{D}_{h, \mathcal{Q}(\alpha, h-1)}$, since the α-HABE dimension takes the minimum of distirbutional eluder dimensions over two distributions. For any $\alpha < q$, consider the function class

$$\mathcal{Q}(\alpha, h-1) = \left\{ Q \in \mathcal{Q}, \left| |\mathbb{E}_{\mu_l(Q)}[Q_l - \mathcal{T}_l Q_{l+1}]| \leqslant \alpha, \ \forall 1 \leqslant l \leqslant h-1 \right. \right\}$$

Now note that $\mathbb{E}_{\mu_l(Q)}[Q_l - \mathcal{T}_l Q_{l+1}] = q\mathbb{1}_{a_1, \dots a_l} - q\mathbb{1}_{a_1^\star, \dots a_l^\star}$. If this is smaller than α, then this is smaller than q and thus must be 0. So, $(a_1, \dots a_{h-1}) = (a_1^\star, \dots a_{h-1}^\star)$ for any $Q \in \mathcal{Q}(\alpha, h-1)$. This also means that any $\phi_h \in \Phi_h$, there is a $Q \in \mathcal{Q}(\alpha, h-1)$ so that

$$\phi_h = Q_h - \mathcal{T}_h Q_{h+1} = q\mathbb{1}_{a_1^\star, \dots a_{h-1}^\star, a_h} - q\mathbb{1}_{a_1^\star, \dots a_{h-1}^\star, a_h^\star}$$

Thus, the size of Φ_h is just A. More importantly, the set $\mathcal{D}_{h, \mathcal{Q}(\alpha, h-1)}$ of distributions $\mu_h(Q)$ induced by $Q \in \mathcal{Q}(\alpha, h-1)$ only include indicators of the form $\mathbb{1}_{a_1^\star, \dots a_{h-1}^\star, a}$ for actions a. Thus, the set of distributions $\mathcal{D}_{\mathcal{Q}(\alpha, h-1)}$ has size A. We know that the distributional eluder dimension $d = \dim_{DE}(\Phi_h, \mathcal{D}_{\mathcal{D}(\alpha, h-1)}, \min(\alpha, Bp/T))$ is bounded by the number of possible distributions $\left|\mathcal{D}_{\mathcal{Q}(\alpha, h-1)}\right|$. So, $d \leqslant A$.

BE dimension: The Bellman differences, from above, are $q\mathbb{1}_{a_1, \dots a_h} - q\mathbb{1}_{a_1^\star, \dots a_h^\star}$. This is an affine transformation of a family of A^H indicator functions. The distributions $\mu_l(Q)$ over trajectories induced by Q include indicators $\mathbb{1}_{a_1', \dots a_l'}$ of all trajectories of length l. Now for any sequence $\mu_1, \dots \mu_n, \mu_{n+1}$ of different indicator distributions not including $a_1^\star, \dots a_l^\star$, we consider the Bellman difference $g_{n+1} = q\mathbb{1}_{a_1, \dots a_h} - q\mathbb{1}_{a_1^\star, \dots a_h^\star}$ with action sequence given by μ_{n+1}. Note that $\mathbb{E}_{\mu_i} g_{n+1} = 0$ for all $i \leqslant n$ but $\mathbb{E}_{\mu_{n+1}} g_{n+1} = q$. This means that the longest possible sequence in the definition of the distributional eluder dimension has length $A^H - 2$. So, the BE dimension is at least $A^H - 2$.

Eluder dimension: The reward function class $\mathcal{F}_h$ is given by all functions of the form $q\mathbb{1}_{a_1, \dots a_h}$. This is a scaled version of a class of A^h indicator functions. Since it contains A^h indicator functions, its eluder dimension is at least A^h.

F.2.2 Sparse intermediate feedback, $\mathcal{H}_p = [H]$

Notice that we get a reward $Ber(q)$ at the *last* step if we took correct sequence of actions $a_1^\star, \dots a_H^\star$, and reward 0 otherwise. It is then easy to see that now, the induced function classes $\mathcal{Q}$ then are given by $\mathcal{Q} = \{(Q_1, \dots Q_H) \mid \exists a_1, \dots a_H \in \mathcal{A} \text{ s.t. } Q_h = q\mathbb{1}_{a_1, \dots a_h}\}$.

α-HABE dimension: This time, note that $\mathbb{E}_{\mu_h(Q)}[Q_l - \mathcal{T}_h Q_{h+1}] = 0$ for all $h \leqslant H - 1$. So, the function class $\Phi_h = \{0\}$ for all $h \leqslant H - 1$. Only for $h = H$ do we have that $\mathbb{E}_{\mu_H(Q)}[Q_H - \mathcal{T}_H Q_{H+1}] = q\mathbb{1}_{a_1, \dots a_H} - q\mathbb{1}_{a_1^\star, \dots a_H^\star}$. Also note that $\mathcal{Q}(\alpha, H-1) = \mathcal{Q}$ for all α since $\mathbb{E}_{\mu_h(Q)}[Q_l - \mathcal{T}_h Q_{h+1}] = 0$ for all $h \leqslant H - 1$. So, this is merely the BE dimension of the problem. Now, the Bellman differences at timestep H are identical to those for the sparse feedback problem, and the distributions $\mathcal{D}_{\mathcal{Q}(\alpha, H-1)} = \mathcal{D}_{\mathcal{Q}}$ since we have established that $\mathcal{Q}(\alpha, H-1) = \mathcal{Q}$. This means that

by the argument for BE dimension in the dense feedback case, we have that the distributional eluder dimension of Φ_H is at least $A^H - 2$, which is then also the α-HABE dimension of this problem.

BE dimension: From the argument for the α-HABE dimension in the sparse case, the BE dimension and the α-HABE dimension match in this case, and are both at least $A^H - 2$.

Eluder dimension: Again, the reward function class $\mathcal{F}_H$ is given by all functions of the form $q\mathbb{1}_{a_1,\dots a_H}$. This is a scaled version of a class of A^H indicator functions. Since it contains A^H indicator functions, its eluder dimension is at least A^H.

□

F.3 Proofs of Lemmas

Recall that $\mathcal{Q}(\alpha, h) = \{Q \in \mathcal{Q} \mid |\mathbb{E}_{\mu_l(Q)}[Q_l - \mathcal{T}_l Q_{l+1}]| \leqslant \alpha,\ \forall 1 \leqslant l \leqslant h\}$, that $\mu_h(Q)$ is the distribution induced on $\tau[h-1], a_h$ by π_Q and $\mathcal{D}_{h,\mathcal{Q}} := \{\mu_h(Q) \mid Q \in \mathcal{Q}\}$.

Lemma 18. *Let $d_h(\varepsilon) := \dim_{DE}(\Phi_h, \mathcal{D}_{h,\mathcal{Q}(\alpha,h-1)}, \varepsilon)$ with*

$$\Phi_h := \left\{Q_h - \mathcal{T}_h Q_{h+1} \middle| Q \in \mathcal{Q}(\alpha, h-1)\right\}$$

Then, we have that for $\beta = c\log(HT\mathcal{N}(\mathcal{Q} \cup \mathcal{G}, 1/T, \|\cdot\|_\infty))$, $\sum_{j=1}^t |\mathbb{E}_{\mu_h(Q^j)}[Q_h^j - \mathcal{T}_h Q_{h+1}^j]|$ is bounded by

$$t\omega + Bp\left(\frac{B^2p^2\beta}{\alpha^2} + 1\right)\left(\sum_{l=1}^{h-1} d_l(\alpha)\right) + \min(d_h(\omega), t)Bp + 2Bp\sqrt{\beta d_h(\omega)t}$$

Proof. We modify the proof of Lemma 41 in [32]. Pick arbitrary h and t and let Ψ_h be the function class given by

$$\begin{aligned}\Phi_h &:= \left\{Q_h - \mathcal{T}_h Q_{h+1} \middle| Q \in \mathcal{Q}(\alpha, h)\right\} \\ &= \left\{Q_h - \mathcal{T}_h Q_{h+1} \middle| (Q_1, \cdots Q_H) \in \mathcal{Q}, |\mathbb{E}_{\mu_l(Q)}[Q_l - \mathcal{T}_l Q_{l+1}]| \leqslant \alpha,\ \forall 1 \leqslant l \leqslant h-1\right\}\end{aligned}$$

Also note that we have the function class Φ_h of timestep h Bellman errors induced by "historically α-accurate" functions - functions whose expected Bellman errors in previous timesteps are smaller than α. The distribution used for computing the expected Bellman errors for previous timesteps is $\mu_l(Q)$.

Now abbreviating $\psi_l^j := Q_l^j - \mathcal{T}_l Q_{l+1}^j$ gives a sequence $\psi_l^1, \dots \psi_l^t$ of functions in Ψ_l for every $1 \leqslant l \leqslant h$. This must have a subsequence $\phi_l^1, \dots \phi_l^{r_l}$ consisting of all the functions in the sequence that lie in Φ_l, for every $1 \leqslant l \leqslant h$. Also let $d_h(\varepsilon) = \dim_{DE}(\Phi_h, \mathcal{D}_h, \varepsilon)$ for any ε. Now note that

$$\begin{aligned}
&\sum_{j=1}^t |\mathbb{E}_{\mu_h(Q^j)}[Q_h^j - \mathcal{T}_h Q_{h+1}^j]| \\
&= \sum_{j=1}^t |\mathbb{E}_{\mu_h(Q^j)}[\psi_h^j]| \\
&\overset{(i)}{=} \sum_{j=1}^t \left|\mathbb{E}_{\mu_h(Q^j)}[\psi_h^j]\right| \mathbb{1}\left(|\mathbb{E}_{\mu_h(Q^j)}[\psi_h^j]| \leqslant \omega\right) \\
&\quad + \sum_{l=1}^{h-1}\sum_{j=1}^t \left|\mathbb{E}_{\mu_h(Q^j)}[\psi_h^j]\right| \mathbb{1}\left(|\mathbb{E}_{\mu_h(Q^j)}[\psi_h^j]| > \omega, Q \in \mathcal{Q}(\alpha, l-1)\backslash\mathcal{Q}(\alpha, l)\right) \\
&\quad + \sum_{j=1}^t \left|\mathbb{E}_{\mu_h(Q^j)}[\psi_h^j]\right| \mathbb{1}\left(|\mathbb{E}_{\mu_h(Q^j)}[\psi_h^j]| > \omega, Q \in \mathcal{Q}(\alpha, h-1)\right) \\
&\leqslant \sum_{j=1}^t \left|\mathbb{E}_{\mu_h(Q^j)}[\psi_h^j]\right| \mathbb{1}\left(|\mathbb{E}_{\mu_h(Q^j)}[\psi_h^j]| \leqslant \omega\right)
\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{l=1}^{h-1}\sum_{j=1}^{t} \left|\mathbb{E}_{\mu_h(Q^j)}[\psi_h^j]\right| \mathbb{1}\left(Q \in \mathcal{Q}(\alpha, l-1) \backslash \mathcal{Q}(\alpha, l)\right) \\
&+ \sum_{j=1}^{t} \left|\mathbb{E}_{\mu_h(Q^j)}[\psi_h^j]\right| \mathbb{1}\left(|\mathbb{E}_{\mu_h(Q^j)}[\psi_h^j]| > \omega, Q \in \mathcal{Q}(\alpha, h-1)\right) \\
\leqslant\ & t\omega + \sum_{l=1}^{h-1}\sum_{j=1}^{t} \left|\mathbb{E}_{\mu_h(Q^j)}[\psi_h^j]\right| \mathbb{1}\left(|\mathbb{E}_{\mu_l(Q^j)}\psi_l^j| > \alpha, Q \in \mathcal{Q}(\alpha, l-1)\right) \\
&\sum_{j=1}^{t} \left|\mathbb{E}_{\mu_h(Q^j)}[\psi_h^j]\right| \mathbb{1}\left(|\mathbb{E}_{\mu_h(Q^j)}[\psi_h^j]| > \omega, Q \in \mathcal{Q}(\alpha, h-1)\right) \\
\overset{(ii)}{\leqslant}\ & t\omega + \sum_{l=1}^{h-1}\sum_{j=1}^{r_l} \left|\mathbb{E}_{\mu_h(Q^j)}[\phi_h^j]\right| \mathbb{1}\left(|\mathbb{E}_{\mu_l(Q^j)}[\phi_l^j]| > \alpha\right) + \sum_{j=1}^{r_h} \left|\mathbb{E}_{\mu_h(Q^j)}[\phi_h^j]\right| \mathbb{1}\left(|\mathbb{E}_{\mu_h(Q^j)}[\phi_h^j]| > \omega\right) \\
\overset{(ii)}{\leqslant}\ & t\omega + Bp\left(\frac{B^2p^2\beta}{\alpha^2}+1\right)\left(\sum_{l=1}^{h-1} d_l(\alpha)\right) + \sum_{j=1}^{r_h} \left|\mathbb{E}_{\mu_h(Q^j)}[\phi_h^j]\right| \mathbb{1}\left(|\mathbb{E}_{\mu_h(Q^j)}[\phi_h^j]| > \omega\right)
\end{aligned}
$$

Here, (i) holds since one of three possibilities holds – either $\left|\mathbb{E}_{\mu_h(Q^j)}[\psi_h^j]\right| \leqslant \omega$, or $|\mathbb{E}_{\mu_h(Q^j)}\psi_l^j| > \omega$ and there is a least $l \leqslant h-1$ so that $Q \in \mathcal{Q}(\alpha, l-1)$ but $Q \notin \mathcal{Q}(\alpha, h-1)$, or $Q \in \mathcal{Q}(\alpha, h-1)$. (ii) holds since if $|\mathbb{E}_{\mu_k(Q^j)}\psi_k^j| \leqslant \alpha$ for all $k \leqslant l-1$, then $\psi_l^j = \phi_l^i$ for some i. Finally, (iii) holds by Proposition 43 of [32] since $\sum_{j=1}^{s-1} \mathbb{E}_{\mu_l(Q^j)}[(\phi_l^j)^2] \leqslant \beta$ by Lemma 39(a) of [32]. While our rewards are stochastic and theirs are not, we can repeat their arguments verbatim after noting that the martingale defined in the beginning of their proof continues to be a martingale even for stochastic rewards that have second moments.

Now arrange the sequence $\left|\mathbb{E}_{\mu_h(Q^j)}\phi_s\right|$ in order to get $e_1, \ldots e_{r_h}$. We can then write

$$
\sum_{j=1}^{t} \left|\mathbb{E}_{\mu_h(Q^j)}[Q_h^j - \mathcal{T}_h Q_{h+1}^j]\right| \leqslant t\omega + Bp\left(\frac{B^2p^2\beta}{\alpha^2}+1\right)\left(\sum_{l=1}^{h-1} d_l(\alpha)\right) + \sum_{j=1}^{r_h} e_j \mathbb{1}(e_j > \omega)
$$

For any $e_j > \omega$, consider arbitrary γ such that $e_j > \gamma > \omega$. This means that by Proposition 43 of [32] again,

$$
j \leqslant \sum_{i=1}^{r_h} \mathbb{1}(e_i > \gamma) \leqslant \left(\frac{B^2p^2\beta}{\gamma^2}+1\right) d_h(\omega)
$$

This means that $\gamma \leqslant Bp\sqrt{\frac{\beta d_h(\omega)}{j - d_h(\omega)}}$ for any such γ. Since $e_j \leqslant Bp$, we get that $e_j \leqslant \min\left(Bp, Bp\sqrt{\frac{\beta d_h(\omega)}{j-d_h(\omega)}}\right)$. Finally, this means that

$$
\begin{aligned}
&\sum_{j=1}^{t} \left|\mathbb{E}_{\mu_h(Q^j)}[Q_h^j - \mathcal{T}_h Q_{h+1}^j]\right| \\
&\leqslant t\omega + Bp\left(\frac{B^2p^2\beta}{\alpha^2}+1\right)\left(\sum_{l=1}^{h-1} d_l(\alpha)\right) + \sum_{j=1}^{r_h} e_j \mathbb{1}(e_j > \omega) \\
&\leqslant t\omega + Bp\left(\frac{B^2p^2\beta}{\alpha^2}+1\right)\left(\sum_{l=1}^{h-1} d_l(\alpha)\right) + \sum_{j=1}^{r_h} \min\left(Bp, Bp\sqrt{\frac{\beta d_h(\omega)}{j - d_h(\omega)}}\right) \\
&\leqslant t\omega + Bp\left(\frac{B^2p^2\beta}{\alpha^2}+1\right)\left(\sum_{l=1}^{h-1} d_l(\alpha)\right) + \min(d_h, r_h)Bp + \sum_{j=1}^{r_h} Bp\sqrt{\frac{\beta d_h(\omega)}{j}} \\
&\leqslant t\omega + Bp\left(\frac{B^2p^2\beta}{\alpha^2}+1\right)\left(\sum_{l=1}^{h-1} d_l(\alpha)\right) + \min(d_h(\omega), r_h)Bp + 2Bp\sqrt{\beta d_h(\omega) r_h}
\end{aligned}
$$

$$\leqslant t\omega + Bp\left(\frac{B^2p^2\beta}{\alpha^2}+1\right)\left(\sum_{l=1}^{h-1} d_l(\alpha)\right) + \min(d_h(\omega),t)Bp + 2Bp\sqrt{\beta d_h(\omega)t}$$

as desired. □

G Proofs for Dueling Feedback

G.1 Proof for Reduction to Confidence-Set Optimism

Theorem 3 (Reduction from Dueling to Confidence-Set-Based Optimism)**.** *If the confidence sets $\mathcal{C}_{\mathcal{M}}(\mathcal{D}_t, \delta)$ satisfy Assumption 2, then the dueling regret $\mathrm{Regret}_D(T)$ of Algorithm 1 is given by*

$$\mathrm{Regret}_D(T) = \widetilde{\mathcal{O}}(C_P(\mathcal{M}, T, \delta) + C_F(\overline{\mathcal{M}}, T, \delta))$$

Remark 5. While the theorem states that we need Assumption 2 from the main paper, we actually use its slightly more refined version – Assumption 3. The less refined version was added to the main paper for brevity.

Proof. For ease of notation, let us use the sets $\mathcal{C}_{\mathcal{M}}(\mathcal{D}_t, \delta)$ given by the pre-image of $\mathcal{C}_{\overline{\mathcal{M}}}(\mathcal{D}_t, \delta)$ under the map $\mathbb{M} \mapsto \overline{\mathbb{M}}$ from Section 4. We first recall that $\mathbb{M}_\star \in \mathcal{C}_{\mathcal{M}}(\mathcal{D}_t, \delta)$ and so $\pi_\star \in \Pi_t$ for all t with probability $1 - \delta/16$. Recall that the value of a duel (π, π') under model $\overline{\mathbb{M}} \leftrightarrow$ is denoted by

$$V_D(\overline{\mathbb{M}}, \pi, \pi') := V(\mathbb{M}, \pi) - V(\mathbb{M}, \pi') = V(\mathbb{P}, f, \pi) - V(\mathbb{P}, g, \pi')$$

We overload notation and use the natural maps $(\mathbb{P}, f) \leftrightarrow \mathbb{M} \mapsto \overline{\mathbb{M}}$ to define

$$V_D(\mathbb{M}, \pi, \pi') := V_D(\overline{\mathbb{M}}, \pi, \pi')$$

For ease of notation, we will then work with $\mathcal{C}_{\mathcal{M}}(\mathcal{D}_t, \delta)$ in this proof until we can. Since $\pi_{i,t} \in \Pi_t$ for $i = 1, 2$, there exists some $\mathbb{M}_{i,t} \in \mathcal{C}_{\mathcal{M}}(\mathcal{D}_t, \delta)$ for $i = 1, 2$ so that $V_D(\mathbb{M}_{i,t}, \pi, \pi_{1,t}) \leqslant 0$ for all π. Note that dueling regret is given below. Inequality (i) is by definition of $\mathbb{M}_{i,t}$, since $V_D(\mathbb{M}_{i,t}, \pi_\star, \pi_{i,t}) \leqslant 0$ for $i = 1, 2$. Inequality (ii) holds by definition of $\pi_{1,t}, \pi_{2,t}$.

$$\begin{aligned}
\mathrm{Regret}_D(T) &= \sum_{t=1}^{T} \sum_{i=1}^{2} V_D(\mathbb{M}_\star, \pi_\star, \pi_{i,t}) \\
&= \sum_{t=1}^{T} \Big[\sum_{i=1}^{2} V_D(\mathbb{M}_\star, \pi_\star, \pi_{i,t}) - V_D(\mathbb{M}_{i,t}, \pi_\star, \pi_{i,t}) + \sum_{i=1}^{2} V_D(\mathbb{M}_{i,t}, \pi_\star, \pi_{i,t}) \Big] \\
&\overset{(i)}{\leqslant} \sum_{i=1}^{2} \sum_{t=1}^{T} \left[V_D(\mathbb{M}_\star, \pi_\star, \pi_{i,t}) - V_D(\mathbb{M}_{i,t}, \pi_\star, \pi_{i,t}) \right] \\
&\overset{(ii)}{\leqslant} \sum_{t=1}^{T} 2 \max_{\mathbb{M}, \mathbb{M}' \in \mathcal{C}_{\mathcal{M}}(\mathcal{D}_t, \delta)} \left[V_D(\mathbb{M}, \pi_{1,t}, \pi_{2,t}) - V_D(\mathbb{M}', \pi_{1,t}, \pi_{2,t}) \right]
\end{aligned}$$

Continuing, we have

$$\begin{aligned}
\mathrm{Regret}_D(T) &\leqslant \sum_{t=1}^{T} 2 \max_{\mathbb{M}, \mathbb{M}' \in \mathcal{C}_{\mathcal{M}}(\mathcal{D}_t, \delta)} \left[V_D(\mathbb{M}, \pi_{1,t}, \pi_{2,t}) - V_D(\mathbb{M}', \pi_{1,t}, \pi_{2,t}) \right] \\
&= 2 \sum_{t=1}^{T} \max_{\mathbb{M}, \mathbb{M}' \in \mathcal{C}_{\mathcal{M}}(\mathcal{D}_t, \delta)} \Big[V_D(\mathbb{M}, \pi_{1,t}, \pi_{2,t}) - V_D(\mathbb{M}_\star, \pi_{1,t}, \pi_{2,t}) + V_D(\mathbb{M}_\star, \pi_{1,t}, \pi_{2,t}) \\
&\qquad\qquad - V_D(\mathbb{M}', \pi_{1,t}, \pi_{2,t}) \Big] \\
&\leqslant 2 \sum_{t=1}^{T} \max_{\mathbb{M}, \mathbb{M}' \in \mathcal{C}_{\mathcal{M}}(\mathcal{D}_t, \delta)} \left[V_D(\mathbb{M}, \pi_{1,t}, \pi_{2,t}) - V_D(\mathbb{M}_\star, \pi_{1,t}, \pi_{2,t}) \right] + \\
&\qquad \max_{\mathbb{M}, \mathbb{M}' \in \mathcal{C}_{\mathcal{M}}(\mathcal{D}_t, \delta)} \left[V_D(\mathbb{M}_\star, \pi_{1,t}, \pi_{2,t}) - V_D(\mathbb{M}', \pi_{1,t}, \pi_{2,t}) \right] \\
&= 2 \sum_{t=1}^{T} \Big[V_D(\widetilde{\mathbb{M}}_t, \pi_{1,t}, \pi_{2,t}) - V_D(\mathbb{M}_\star, \pi_{1,t}, \pi_{2,t}) \Big] + \Big[V_D(\mathbb{M}_\star, \pi_{1,t}, \pi_{2,t}) - V_D(\widetilde{\mathbb{M}}'_t, \pi_{1,t}, \pi_{2,t}) \Big]
\end{aligned}$$

where $\widetilde{\mathbb{M}}_t$ and $\widetilde{\mathbb{M}}'_t$ are the respective maximisers. It suffices to analyse only one of the terms, as a consequence of the symmetry of Assumption 3.

We can now use the fact that $\mathbb{M}$ is described by $(\mathbb{P}, f)$ to analyse the first term, letting $\widetilde{\mathbb{M}}_t \leftrightarrow (\widetilde{\mathbb{P}}_t, \widetilde{f}^t)$.

$$
\begin{aligned}
&\sum_{t=1}^{T}\Big[V_D(\widetilde{\mathbb{M}}_t, \pi_{1,t}, \pi_{2,t}) - V_D(\mathbb{M}_\star, \pi_{1,t}, \pi_{2,t})\Big] \\
&= 2\sum_{t=1}^{T}\Big[V_D(\widetilde{\mathbb{P}}_t, \widetilde{f}^t, \pi_{1,t}, \pi_{2,t}) - V_D(\mathbb{P}_\star, f^\star, \pi_{1,t}, \pi_{2,t})\Big] \\
&\leqslant 2\sum_{t=1}^{T}\Big[V_D(\widetilde{\mathbb{P}}_t, \widetilde{f}^t, \pi_{1,t}, \pi_{2,t}) - V_D(\mathbb{P}_\star, \widetilde{f}^t, \pi_{1,t}, \pi_{2,t})\Big] + \Big[V_D(\mathbb{P}_\star, \widetilde{f}^t, \pi_{1,t}, \pi_{2,t}) - V_D(\mathbb{P}_\star, f^\star, \pi_{1,t}, \pi_{2,t})\Big] \\
&\leqslant 2\sum_{t=1}^{T}\Big[V_D(\widetilde{\mathbb{P}}_t, \widetilde{f}^t, \pi_{1,t}, \pi_{2,t}) - V_D(\mathbb{P}_\star, \widetilde{f}^t, \pi_{1,t}, \pi_{2,t})\Big] + \Big[V_D(\mathbb{P}_\star, \widetilde{f}^t, \pi_{1,t}, \pi_{2,t}) - V_D(\mathbb{P}_\star, f^\star, \pi_{1,t}, \pi_{2,t})\Big] \\
&\overset{(i)}{=} 2\sum_{t=1}^{T}\Big[V(\widetilde{\mathbb{P}}_t, \widetilde{f}^t, \pi_{1,t}) - V(\mathbb{P}_\star, \widetilde{f}^t, \pi_{1,t})\Big] - \Big[V(\widetilde{\mathbb{P}}_t, \widetilde{f}^t, \pi_{2,t}) - V(\mathbb{P}_\star, \widetilde{f}^t, \pi_{2,t})\Big] \\
&\qquad + \Big[V_D(\mathbb{P}_\star, \widetilde{f}^t, \pi_{1,t}, \pi_{2,t}) - V_D(\mathbb{P}_\star, f^\star, \pi_{1,t}, \pi_{2,t})\Big] \\
&\overset{(ii)}{=} 2\sum_{t=1}^{T}\Big[V(\widetilde{\mathbb{P}}_t, \widetilde{f}^t, \pi_{1,t}) - V(\mathbb{P}_\star, \widetilde{f}^t, \pi_{1,t})\Big] - \Big[V(\widetilde{\mathbb{P}}_t, \widetilde{f}^t, \pi_{2,t}) - V(\mathbb{P}_\star, \widetilde{f}^t, \pi_{2,t})\Big] \\
&\qquad + \Big[V(\mathbb{P}_\star \otimes \mathbb{P}_\star, \overline{f}^t, (\pi_{1,t}, \pi_{2,t})) - V(\mathbb{P}_\star \otimes \mathbb{P}_\star, \overline{f}^\star, (\pi_{1,t}, \pi_{2,t}))\Big]
\end{aligned}
$$

Where (i) holds by the definition of V_D and V, and (ii) holds in the product MDP $\overline{\mathbb{M}}_\star$ once we define $\overline{f}^t_h((\tau_1, \tau_2)[h]) := \widetilde{f}^t_h(\tau_1[h]) - \widetilde{f}^t_h(\tau_2[h])$ and recall that $\overline{\mathbb{P}}_\star = \mathbb{P}_\star \otimes \mathbb{P}_\star$. Now, we can immediately apply Assumption 3 to the last line in two different ways. For the first two terms, we apply the first point in the assumption to each under cardinal feedback for MDP $\mathbb{M}_\star$, noting that the datasets $\mathcal{D}_t$ contain trajectories from $\pi_{1,t}$ as well as $\pi_{2,t}$. For the last term, we apply the second point in the assumption under cardinal feedback for the MDP $(\overline{\mathbb{P}}_\star, \overline{f}^\star)$.

This gives us that with probability $1 - \delta$,

$$\mathrm{Regret}(T) = \widetilde{\mathcal{O}}(C_P(\mathcal{M}, T, \delta) + C_F(\overline{\mathcal{M}}, T, \delta))$$

□

We have the following lemma, which is an immediate consequence of

Lemma 2 (Relating $\mathcal{F}$ and $\overline{\mathcal{F}}$)**.** *For any function class $\mathcal{F}$,* $\dim_E(\overline{\mathcal{F}}, \varepsilon) \leqslant 9\dim_E(\mathcal{F}, \varepsilon/2)$.

Proof. Let $\overline{d}_h = \dim_E(\overline{\mathcal{F}}_h, \varepsilon)$. Pick the ε' so that there is a sequence of $\overline{d}_h$ pairs $\overline{\tau}_j, j = 1 \to \overline{d}_h$ of length h trajectories, where each one is ε'-independent of its predecessors. Note that $\overline{\tau}_j = (\tau_{1,j}, \tau_{2,j})$. We now inductively build a sequence i_j so that each $\tau_{i_j,j}$ is $\varepsilon'/2$-independent of its predecessors.

Pick the first i_1 arbitrarily. Now assume that we have built the sequence until index $j = k$. Also, by definition of this sequence, there exist $\overline{f}_j, \overline{f}'_j$, we have $\sqrt{\sum_{j=1}^{k}(\overline{f}_j(\overline{\tau}_j) - \overline{f}'_j(\overline{\tau}_j))^2} \leqslant \varepsilon'$ but $|\overline{f}_{k+1}(\overline{\tau}_j) - \overline{f}'_{k+1}(\overline{\tau}_j)| \geqslant \varepsilon'$. Since $a^2 + b^2 \leqslant 2(a+b)^2$, we have that

$$
\begin{aligned}
\sqrt{\sum_{j=1}^{k}(f_j(\tau_{i_j,j}) - f'_j(\tau_{i_j,j}))^2} &\leqslant \sqrt{\sum_{j=1}^{k}(f_j(\tau_{i_j,j}) - f'_j(\tau_{i_j,j}))^2 + (f_j(\tau_{3-i_j,j}) - f'_j(\tau_{3-i_j,j}))^2} \\
&\leqslant \sqrt{\sum_{j=1}^{k} 2(\overline{f}_j(\overline{\tau}_j) - \overline{f}'_j(\overline{\tau}_j))^2} \leqslant \sqrt{2}\varepsilon'
\end{aligned}
$$

Additionally, since

$$|f_{k+1}(\tau_{1,k+1}) - f'_{k+1}(\tau_{1,k+1})| + |f_{k+1}(\tau_{2,k+1}) - f'_{k+1}(\tau_{2,k+1})| \geqslant |\overline{f}_{k+1}(\overline{\tau}_j) - \overline{f}'_{k+1}(\overline{\tau}_j)| \geqslant \varepsilon'$$

. So, there is an i_{k+1} so that

$$|f_{k+1}(\tau_{i_{k+1},k+1}) - f'_{k+1}(\tau_{i_{k+1},k+1})| \geqslant \varepsilon'/2$$

So, we have a sequence $x_j := \tau_{i_j,j}$ and a sequence of pairs of functions f_j, f'_j so that for any $1 \leqslant k \leqslant \overline{d}_h$, $\sum_{j=1}^{k}(f_j(x_j) - f'_j(x_j))^2 \leqslant 2(\varepsilon')^2$ but $|f_{k+1}(x_{k+1}) - f'_{k+1}(x_{k+1})| \geqslant \varepsilon'/2$. This implies the following. Inequality (i) holds by Proposition 43 of [32] upon setting $\beta = 2(\varepsilon')^2$ and setting the proposition's ε to $\varepsilon'/2$. Inequality (ii) holds since $\varepsilon'/2 \geqslant \varepsilon/2$.

$$\begin{aligned}
\overline{d}_h &= \sum_{j=1}^{\overline{d}_h} \mathbb{1}(|f_j(x_j) - f'_j(x_j)| \geqslant \varepsilon'/2) \\
&\overset{(i)}{\leqslant} \left(\frac{2(\varepsilon')^2}{(\varepsilon'/2)^2} + 1\right) \dim_E(\mathcal{F}_h, \varepsilon/2) \\
&= 9 \dim_E(\mathcal{F}_h, \varepsilon'/2) \\
&\leqslant 9 \dim_E(\mathcal{F}_h, \varepsilon/2)
\end{aligned}$$

This establishes our claim. □

We have the following immediate corollary of Theorem 4, Theorem 7 and Lemma 2.

Corollary 1 (Dueling Regret using POR-UCRL Confidence Sets)**.** *The confidence sets from POR-UCRL satisfy Assumption 2 and using them in Algorithm 1 leads to the following regret bound* $\mathrm{Regret}_D(T) = \widetilde{\mathcal{O}}\left(\left(pS\sqrt{HA} + \sum_{h\in\mathcal{H}_p}\sqrt{d_{E,h}d_{C,h}}\right)\sqrt{T}\right)$.

G.2 Reduction to Bonus-Based Optimism

We define the reduction using the algorithm below.

Algorithm 8 Reduction from Dueling to Cardinal Bonus-Based Optimism

1: **Input** Known reward function $\{r_h\}_{h=1}^H$, method $\texttt{Est}(\mathcal{D})$ to estimate $\hat{\mathbb{P}}_{\mathcal{D}}$ and $\overline{f}_{\mathcal{D}}$ from dataset $\mathcal{D}$, bonus functions $b^{\mathcal{D}}_{\overline{\mathcal{F}}}(\mathbb{P}, \pi, \delta)$ and $b^{\mathcal{D}}_{\mathcal{P}}(\mathbb{P}, \pi, \delta)$, confidence level δ.
2: Initialize dataset $\mathcal{D}_1 \leftarrow \{\}$
3: **for** $t = 1, ..., T$ **do**
4: Compute good set Π_t {Valid $\pi_\star$ candidates}

$$\Pi_t := \Big\{\pi \in \Pi \Big| V_D((\hat{\mathbb{P}}_{\mathcal{D}_t}, \overline{f}_{\mathcal{D}_t}), \pi, \pi_1) + b_{\overline{\mathcal{F}}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, (\pi, \pi_1), \delta) + z(Bp) b_{\mathcal{P}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi, \delta) + z(Bp) b_{\mathcal{P}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_1, \delta) \geqslant 0,\ \forall \pi_1 \in \Pi\Big\}$$

5: Pick $(\pi_{1,t}, \pi_{2,t})$ given by {Most uncertain duel}

$$\arg\max_{\pi, \pi' \in \Pi_t} b_{\overline{\mathcal{F}}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, (\pi, \pi'), \delta) + z(Bp) b_{\mathcal{P}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi, \delta) + z(Bp) b_{\mathcal{P}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_1, \delta)$$

6: Collect trajectories $\tau_{t,i} = \Big\{(s^t_{h,i}, a^t_{h,i}))\Big\}_{h=1}^H$ along with feedback $\{o_h\}_{h \in \mathcal{H}_p}$ by sampling from $\mathbb{P}_\star^{\pi_{i,t}}$ for $i = 1, 2$.
7: Append the data to $\mathcal{D}_t$ to get $\mathcal{D}_{t+1}$, update estimates and bonuses.
8: **end for**

Theorem 10 (Reduction from Dueling to Bonus-Based Optimism)**.** *If the bonuses and estimates used in Algorithm 8 satisfy Assumption 3, then with probability* $1 - \delta$*, the dueling regret* $\mathrm{Regret}_D(T)$ *of Algorithm 8 is given by*

$$\mathrm{Regret}_D(T) = \widetilde{\mathcal{O}}(C_P(\mathcal{M}, T, \delta) + C_F(\overline{\mathcal{M}}, T, \delta))$$

Proof. Recall that the value of a duel (π, π') under model $\overline{\mathbb{M}} \leftrightarrow \mathbb{M} \leftrightarrow (\mathbb{P}, f)$ is denoted by

$$V_D(\overline{\mathbb{M}}, \pi, \pi') := V(\mathbb{M}, \pi) - V(\mathbb{M}, \pi') = V(\mathbb{P}, f, \pi) - V(\mathbb{P}, g, \pi')$$

We overload notation and use the natural bijection $\mathcal{M} \leftrightarrow \overline{\mathcal{M}}$ to define

$$V_D(\mathbb{M}, \pi, \pi') := V_D(\overline{\mathbb{M}}, \pi, \pi')$$

For ease of notation in the proof, we often work with an arbitrary pre-image $\hat{f}_{\mathcal{D}}$ of $\overline{f}_{\mathcal{D}}$ under the map $f \mapsto \overline{f}$. A careful read will confirm that this does not affect the correctness of any of the statements. First note that $\pi_\star \in \Pi_t$ for all T with probability $1 - \delta/16$ since the following hold uniformly over all $\pi_1 \in \Pi$

$$\begin{aligned}
-V_D((\hat{\mathbb{P}}_{\mathcal{D}_t}, \hat{f}_{\mathcal{D}_t}), \pi_\star, \pi_1) &= V(\hat{\mathbb{P}}_{\mathcal{D}_t}, \hat{f}_{\mathcal{D}_t}, \pi_1) - V(\hat{\mathbb{P}}_{\mathcal{D}_t}, \hat{f}_{\mathcal{D}_t}, \pi_\star) \\
&= \Big[V(\hat{\mathbb{P}}_{\mathcal{D}_t}, \hat{f}_{\mathcal{D}_t}, \pi_1) - V(\mathbb{P}_\star, \hat{f}_{\mathcal{D}_t}, \pi_1)\Big] - \Big[V(\mathbb{P}_\star, \hat{f}_{\mathcal{D}_t}, \pi_1) - V(\hat{\mathbb{P}}_{\mathcal{D}_t}, \hat{f}_{\mathcal{D}_t}, \pi_1)\Big] \\
&\quad + V(\mathbb{P}_\star, f^\star, \pi_1) - V(\mathbb{P}_\star, f^\star, \pi_\star) \\
&\quad + V_D((\mathbb{P}_\star, \overline{f}_{\mathcal{D}_t}), \pi_1, \pi_\star) - V_D((\mathbb{P}_\star, f^\star), \pi_1, \pi_\star) \\
&\leqslant z(Bp) b_{\mathcal{P}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_\star, \delta) + z(Bp) b_{\mathcal{P}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_1, \delta) \\
&\quad + 0 \\
&\quad + b_{\overline{\mathcal{F}}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, (\pi_\star, \pi_1), \delta) + \\
&= b_{\overline{\mathcal{F}}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, (\pi_\star, \pi_1), \delta) + z(Bp) b_{\mathcal{P}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_\star, \delta) + z(Bp) b_{\mathcal{P}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_1, \delta)
\end{aligned}$$

where the inequality holds by Assumption 4 and the optimality of $\pi_\star$ in the true model. Note let $\hat{\mathbb{M}}_t$ be the model given by $\hat{\mathbb{P}}_{\mathcal{D}_t}, \hat{f}_{\mathcal{D}_t}$ and let $\overline{\mathbb{M}}_t$ be the corresponding model in $\overline{\mathcal{M}}$. We make the following abbreviation:

$$b_{\overline{\mathcal{M}}}(\overline{\mathbb{M}}_t, (\pi, \pi'), \delta) := b_{\overline{\mathcal{F}}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, (\pi, \pi'), \delta) + z(Bp) b_{\mathcal{P}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi, \delta) + z(Bp) b_{\mathcal{P}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi', \delta)$$

$$
\begin{aligned}
\mathrm{Regret}_D(T) &= \sum_{t=1}^{T}\sum_{i=1}^{2} V_D(\mathbb{M}_\star, \pi_\star, \pi_{i,t}) \\
&= \sum_{t=1}^{T}\Big[\sum_{i=1}^{2} V_D(\mathbb{M}_\star, \pi_\star, \pi_{i,t}) - V_D(\hat{\mathbb{M}}_t, \pi_\star, \pi_{i,t}) + b_{\overline{\mathcal{M}}}(\overline{\mathbb{M}}_t, (\pi_\star, \pi_{i,t}), \delta) \\
&\qquad + \sum_{i=1}^{2} V_D(\hat{\mathbb{M}}_t, \pi_\star, \pi_{i,t}) - b_{\overline{\mathcal{M}}}(\overline{\mathbb{M}}_t, (\pi_\star, \pi_{i,t}), \delta)\Big] \\
&\overset{(i)}{\leqslant} \sum_{i=1}^{2}\sum_{t=1}^{T}\Big[V_D(\mathbb{M}_\star, \pi_\star, \pi_{i,t}) - V_D(\hat{\mathbb{M}}_t, \pi_\star, \pi_{i,t}) + b_{\overline{\mathcal{M}}}(\overline{\mathbb{M}}_t, (\pi_\star, \pi_{i,t}), \delta)\Big]
\end{aligned}
$$

Inequality (i) holds since $V_D(\hat{\mathbb{M}}_t, \pi_\star, \pi_{i,t}) = -V_D(\hat{\mathbb{M}}_t, \pi_{i,t}, \pi_\star)$, and $\pi_{i,t} \in \Pi_t$ for $i = 1, 2$ implies that

$$
V_D(\hat{\mathbb{M}}_t, \pi_{i,t}, \pi_\star) + b_{\overline{\mathcal{F}}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, (\pi_\star, \pi_1), \delta) + z(Bp) b_{\mathcal{P}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_\star, \delta) + z(Bp) b_{\mathcal{P}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_1, \delta) \geqslant 0
$$

Now note that the following holds uniformly over all timesteps t with probability $1 - 3\delta/8$ for $i = 1, 2$ simultaneously using Assumption 4 multiple times and applying a union bound.

$$
\begin{aligned}
V_D(\mathbb{M}_\star, \pi_\star, \pi_{i,t}) - V_D(\hat{\mathbb{M}}_t, \pi_\star, \pi_{i,t}) &= V_D((\mathbb{P}_\star, \overline{f}^\star), \pi_\star, \pi_{i,t}) - V_D((\hat{\mathbb{P}}_{\mathcal{D}_t}, \overline{f}_{\mathcal{D}_t}), \pi_\star, \pi_{i,t}) \\
&= V_D((\mathbb{P}_\star, \overline{f}^\star), \pi_\star, \pi_{i,t}) - V_D((\hat{\mathbb{P}}_{\mathcal{D}_t}, \overline{f}^\star), \pi_\star, \pi_{i,t}) \\
&\qquad + V_D((\hat{\mathbb{P}}_{\mathcal{D}_t}, \overline{f}^\star), \pi_\star, \pi_{i,t}) - V_D((\hat{\mathbb{P}}_{\mathcal{D}_t}, \overline{f}_{\mathcal{D}_t}), \pi_\star, \pi_{i,t}) \\
&= V(\mathbb{P}_\star, \overline{f}^\star, \pi_\star) - V(\hat{\mathbb{P}}_{\mathcal{D}_t}, \overline{f}^\star, \pi_\star) + V(\hat{\mathbb{P}}_{\mathcal{D}_t}, \overline{f}^\star, \pi_{i,t}) - V(\mathbb{P}_\star, \overline{f}^\star, \pi_{i,t}) \\
&\qquad + V_D((\hat{\mathbb{P}}_{\mathcal{D}_t}, \overline{f}^\star), \pi_\star, \pi_{i,t}) - V_D((\hat{\mathbb{P}}_{\mathcal{D}_t}, \overline{f}_{\mathcal{D}_t}), \pi_\star, \pi_{i,t}) \\
&= z(Bp) b_{\mathcal{P}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_\star, \delta) + z(Bp) b_{\mathcal{P}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_{i,t}, \delta) \\
&\qquad + b_{\overline{\mathcal{F}}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, (\pi_\star, \pi_{i,t}), \delta) \\
&= b_{\overline{\mathcal{M}}}(\overline{\mathbb{M}}_t, (\pi_\star, \pi_{i,t}), \delta)
\end{aligned}
$$

So, with probability $1 - 3\delta/16$, we have that

$$
\begin{aligned}
\mathrm{Regret}_D(T) &\leqslant \sum_{t=1}^{T}\sum_{i=1}^{2} b_{\overline{\mathcal{M}}}(\overline{\mathbb{M}}_t, (\pi_\star, \pi_{i,t}), \delta) \\
&\overset{(i)}{\leqslant} 2\sum_{t=1}^{T} b_{\overline{\mathcal{M}}}(\overline{\mathbb{M}}_t, (\pi_{1,t}, \pi_{2,t}), \delta) \\
&= 2\sum_{t=1}^{T} z(Bp) b_{\mathcal{P}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_{1,t}, \delta) + z(Bp) b_{\mathcal{P}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, \pi_{2,t}, \delta) + b_{\overline{\mathcal{F}}}(\hat{\mathbb{P}}_{\mathcal{D}_t}, (\pi_{1,t}, \pi_{2,t}), \delta) \\
&\overset{(ii)}{\leqslant} \widetilde{\mathcal{O}}\left(\sum_{t=1}^{T}(z_1(Bp) b_{\mathcal{P}}(\mathbb{P}_\star, \pi_{1,t}, \delta) + z_1(Bp) b_{\mathcal{P}}(\mathbb{P}_\star, \pi_{2,t}, \delta) + b_{\overline{\mathcal{F}}}(\mathbb{P}_\star, (\pi_{1,t}, \pi_{2,t}), \delta)\right)
\end{aligned}
$$

where inequality (i) holds since $(\pi_{1,t}, \pi_{2,t}) = \arg\max_{\pi, \pi' \in \Pi_t} b_{\overline{\mathcal{M}}}(\overline{\mathbb{M}}_t, (\pi, \pi'), \delta)$ and inequality (ii) holds with probability $1 - 3\delta/8$ by 6 applications of the change of measure inequality in Assumption 4.

Now, we can use the fact that Assumption 4 is satisfied again to conclude that with probability $1 - \delta/32$.

$$
\sum_{t=1}^{T}(z_1(Bp) b_{\mathcal{P}}(\mathbb{P}_\star, \pi_{1,t}, \delta) + z_1(Bp) b_{\mathcal{P}}(\mathbb{P}_\star, \pi_{2,t}, \delta) + b_{\overline{\mathcal{F}}}(\mathbb{P}_\star, (\pi_{1,t}, \pi_{2,t}), \delta) = \widetilde{\mathcal{O}}(C_P(\mathcal{M}, T, \delta) + C_F(\overline{\mathcal{M}}, T, \delta))
$$

Taking a union bound over all inequalities stated so far, we have the following with probability $1 - \delta$

$$
\mathrm{Regret}_D(T) = \widetilde{\mathcal{O}}(C_P(\mathcal{M}, T, \delta) + C_F(\overline{\mathcal{M}}, T, \delta))
$$

as desired. □

Again, the following corollary is immediate from Theorem 5, Theorem 9 and Lemma 2.

Corollary 7. *By using POR-UCBVI as the algorithm in the dueling reduction in Algorithm 8, we can get a bound on the dueling regret given by*

$$\mathrm{Regret}_D(T) = \widetilde{\mathcal{O}}\left(\left(pC(H,S,A) + \sum_{h\in\mathcal{H}_p}\sqrt{\bar{d}_{E,h}(d_{C,h}+H)}\right)\sqrt{T}\right)$$

where $\bar{d}_{E,h} = \dim_E\left(\mathcal{F}_h, \frac{B}{2T}\right)$.