Model-Free Robust φ-Divergence Reinforcement Learning Using Both Offline and Online Data

Kishan Panaganti [1] **Adam Wierman** [1] **Eric Mazumdar** [1]

Abstract

The robust φ-regularized Markov Decision Process (RRMDP) framework focuses on designing control policies that are robust against parameter uncertainties due to mismatches between the simulator (nominal) model and real-world settings. This work makes *two* important contributions. First, we propose a *model-free* algorithm called *Robust φ-regularized fitted Q-iteration* for learning an ε-optimal robust policy that uses only the historical data collected by rolling out a behavior policy (with *robust exploratory* requirement) on the nominal model. To the best of our knowledge, we provide the *first* unified analysis for a class of φ-divergences achieving robust optimal policies in high-dimensional systems of arbitrary large state space with general function approximation. Second, we introduce the *hybrid robust φ-regularized reinforcement learning* framework to learn an optimal robust policy using both historical data and online sampling. Towards this framework, we propose a model-free algorithm called *Hybrid robust Total-variation-regularized Q-iteration*. To the best of our knowledge, we provide the *first* improved out-of-data-distribution assumption in large-scale problems of arbitrary large state space with general function approximation under the hybrid robust φ-regularized reinforcement learning framework.

1. Introduction

Online Reinforcement Learning (RL) agents learn through online interactions and exploration in environments and have been shown to perform well in structured domains such as Chess and Go (Silver et al., 2018), fast chip placements in semiconductors (Mirhoseini et al., 2021), fast transform computations in mathematics (Fawzi et al., 2022), and more. However, online RL agents (Botvinick et al., 2019) are known to suffer sample inefficiency due to complex exploration strategies in sophisticated environments. To overcome this, learning from available historical data has been studied using offline RL protocols (Levine et al., 2020). However, offline RL agents suffer from out-of-data-distribution (Yang et al., 2021; Robey et al., 2020) due to the lack of online exploration. Recent work Song et al. (2023) proposes another learning setting called *hybrid RL* that makes the best of both offline and online RL worlds. In particular, hybrid RL agents have access to both offline data (to reduce exploration overhead) and online interaction with the environment (to mitigate the out-of-data-distribution issue).

All three of these approaches (online, offline, and hybrid RL) require training environments (simulators) that closely represent real-world environments. However, time-varying real-world environments (Maraun, 2016), sensor degradations (Chen et al., 1996), and other adversarial disturbances in practice (Pioch et al., 2009) mean that even high-fidelity simulators are not enough (Schmidt et al., 2015; Shah et al., 2018). RL agents are known to fail due to these mismatches between training and testing environments (Sünderhauf et al., 2018; Lesort et al., 2020). As a result, robust RL (Mankowitz et al., 2020; Panaganti & Kalathil, 2021a) has received increasing attention due to the potential for it to alleviate the issue of mismatches between the simulator and real-world environments.

Robust RL agents are built using the robust Markov Decision Process (RMDP) (Iyengar, 2005; Nilim & El Ghaoui, 2005) framework. In this framework, the goal is to find an optimal policy that is robust, i.e., performs uniformly well across a set of models (transition probability functions). This is formulated via a max-min problem, and the set of models is typically constructed around a simulator model (transition probability function) with some notion of divergence or distance function. We refer to the simulator model as any *nominal model* that is provided to RL agents.

The RMDP framework in RL is identical to the Distributionally Robust Optimization (DRO) framework in supervised learning (Duchi & Namkoong, 2018; Chen et al., 2020).

[1]Computing + Mathematical Sciences Department, California Institute of Technology. Correspondence to: Kishan Panaganti <kpb@caltech.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Similar to RMDP, DRO is a min-max problem aiming to minimize a loss function uniformly over the set of distributions constructed around the training distribution of the input space. However, developing model-free algorithms for DRO problems with general φ-divergences (see Equation (1)) is known to be hard (Namkoong & Duchi, 2016) due to their inherent non-linear and multi-level optimization structure. Additionally, developing model-free robust RL agents is also challenging (Iyengar, 2005; Duchi & Namkoong, 2018) for high-dimensional sequential decision-making systems under general function approximation.

To overcome this issue, in this work, we develop robust RL agents for the RRMDP framework, which is an equivalent alternative form of RMDP. A natural φ-divergence regularization extension to the problem of RMDP gives way for this new RRMDP framework introduced in Yang et al. (2023); Zhang et al. (2024), under different names. It is built upon the penalized DRO problem (Levy et al., 2020; Jin et al., 2021b), that is, the φ-divergence regularization version of the DRO problem. In particular, we focus on developing an **offline robust RL algorithm** for a class of φ-divergences under the RRMDP framework with arbitrarily large state spaces, using only offline data with general function approximation. Towards this, as the *first main contribution*, we propose the *Robust φ-regularized fitted Q-iteration* (RPQ) model-free algorithm and provide its performance guarantee for a class of φ-divergences with a unified analysis. We refer to algorithms as *model-free* if they do not explicitly estimate the underlying nominal model. We address the following important (suboptimality and sample complexity) questions: *What is the rate of suboptimality gap achieved between the optimal robust value and the value of RPQ policy? How many offline data samples from the nominal model are required to learn an ε-optimal robust policy?* We discuss challenges and present these results in Section 2.

In this work, we also develop and study a novel **hybrid robust RL algorithm** under the RRMDP framework using both offline data and online interactions with the nominal model. We make this *second main contribution* to this work since hybrid RL overcomes the out-of-data-distribution issue in offline RL. Towards this, we propose the *Hybrid robust Total-variation-regularized Q-iteration* (HyTQ: pronounced *height-Q*) algorithm and provide its performance guarantee under improved assumptions. Notably, the offline data-generating distribution must only cover the distribution that the optimal robust policy samples out on the nominal model, whereas before we needed it to cover any distribution uniformly. This is how online interactions help mitigate the out-of-data-distribution issue of offline RL and offline robust RL. We now address the cumulative suboptimality question in addition to sample complexity: *What is the rate of cumulative suboptimality gap achieved between the optimal robust value and the value of HyTQ iteration policies?* We discuss challenges and present these results in Section 3.

Related Work. Among all the previous works that provide model-free methods, here we only mention the ones closest to ours. We discuss more related works in Appendix A. Panaganti et al. (2022) proposed a Q-iteration offline robust RL algorithm in the RMDP framework only for the total variation φ-divergence. Bruns-Smith & Zhou (2023) proposed a Q-iteration offline robust RL algorithm in the RMDP framework to solve causal inference under unobserved confounders. Zhou et al. (2023) proposed an actor-critic robust RL algorithm in RMDP for integral probability metric. Zhang et al. (2024) proposed a Q-iteration offline robust RL algorithm in the RRMDP framework only for the Kullback-Leibler φ-divergence. Blanchet et al. (2023) proposed specialized robust RL algorithms for the total variation and Kullback-Leibler φ-divergences offering unified analyses for linear, kernels, and factored nominal models under the finite state-action setting. Other line of work (Liu et al., 2022; Liang et al., 2023; Wang et al., 2023a;b; Yang et al., 2023) provide model-free robust RL algorithms based on classical Q-learning methods in finite state-action spaces. We provide more insightful comparisons in Table 1. *To the best of our knowledge, this is the first work that addresses a wide class of robust RL problems (like the general φ-divergence) with arbitrary large state space using general function approximation under mild assumptions (like the robust Bellman error transfer coefficient).*

Notation. We use the equality sign (=) for pointwise equality in vectors and matrices. For any $x \in \mathbb{R}$, let $(x)_+ = \max\{x, 0\}$. For any vector x and positive semidefinite matrix A, the squared matrix norm is $\|x\|_A^2 = x^\top A x$. The set of probability distributions over $\mathcal{X}$, with cardinality $|\mathcal{X}|$, is denoted as $\Delta(\mathcal{X})$, and its power set sigma algebra as $\Sigma(\mathcal{X})$. For any function f that takes (s, a, r, s') as input, define the expectation w.r.t. the dataset $\mathcal{D}$ (or empirical expectation) as $\mathbb{E}_{\mathcal{D}}[f(s_i, a_i, r_i, s_i')] = \frac{1}{N}\sum_{(s_i,a_i,r_i,s_i')\in\mathcal{D}} f(s_i, a_i, r_i, s_i')$. For any positive integer H, set $[H]$ denotes $\{0, 1, \cdots, H-1\}$. Define ℓ_2 and ℓ_1 norms as $\|x\|_{2,\mu} = \sqrt{\mathbb{E}_\mu[x^2]}$ and $\|x\|_{1,\mu} = \mathbb{E}_\mu[|x|]$. $p \ll q$ denotes a probability distribution p is absolutely continuous w.r.t a probability distribution q. We use $\mathcal{O}(\cdot)$ to ignore universal constants less than 300 and $\tilde{\mathcal{O}}(\cdot)$ to ignore universal constants less than 300 and the polylog terms depending on problem parameters.

2. Offline Robust φ-Regularized Reinforcement Learning

We start with preliminaries and the problem formulation.

Infinite-Horizon Markov Decision Process: An infinite-horizon discounted Markov Decision Process (γMDP) is a tuple $(\mathcal{S}, \mathcal{A}, R, P, \gamma, d_0)$ where $\mathcal{S}$ is a countably large state-space, $\mathcal{A}$ is a finite set of actions, $R : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is

Algorithm	Algorithm-type	Data Coverage	Dataset Type	Robust	Suboptimality
(Panaganti et al., 2022, Alg.1)	FQI	all-policy	offline	TV	$\frac{V_{\max}^3\sqrt{\log(\lvert\mathcal{F}\rvert\lvert\mathcal{G}\rvert)}}{\rho N^{1/2}}$
(Zhang et al., 2024, Alg.1)	FQI	all-policy*	offline	KL	$\frac{\lambda V_{\max}^2\sqrt{\log(\lvert\mathcal{F}\rvert)}}{e^{-V_{\max}/\lambda}N^{1/2}}$
(Yang et al., 2023, Alg.2)	QL	uniform-policy	offline Markov	φ	$\frac{V_{\max}^3\sqrt{\log(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert)}}{d_{\min}^3 c(\lambda)N^{1/3}}$
RPQ (**ours:** Algorithm 1)	FQI	all-policy*	offline	φ	$\frac{V_{\max}^3\sqrt{\log(\lvert\mathcal{F}\rvert\lvert\mathcal{G}\rvert)}}{c(\lambda)N^{1/2}}$
HyTQ (**ours:** Algorithm 2)†	FQI	single-policy	offline + online non-Markov	TV	$\frac{V_{\max}(\lambda+V_{\max})\log(\lvert\mathcal{F}\rvert\lvert\mathcal{G}\rvert)}{N^{1/2}}$

Table 1. Comparison of model-free φ-divergence robust RL algorithms. In the *algorithm-type* column, Fitted Q-Iteration (FQI) uses least-squares regression and Q-Learning (QL) uses stochastic approximation updates. In the *data coverage* column, *uniform-policy* stipulates a data-generating policy to cover the entire state-action space. *all-policy* is where the data-generating policy should cover the state-action space covered by all non-stationary policies, and *single-policy* is where it covers the state-action space covered by the optimal robust policy, on the nominal model. * denotes the coverage should include all the models in robust sets designed by the divergences in the *robust* column. The *dataset type* column mentions the type of dataset collected with a data-generating policy for training corresponding algorithms where *offline* indicates i.i.d. historical dataset on the nominal model, *offline Markov* indicates Markovian dataset induced on the nominal model, and *online non-Markov* indicates a history dependent dataset as a collection of Markovian datasets induced on the nominal model by a set of learned policies. Finally, the suboptimality column is the statistical upper bound for the difference between the optimal robust value and the robust value achieved by the algorithm. Here $V_{\max}$ is either H or $(1-\gamma)^{-1}$ effective horizon factors. ρ is the robustness radius parameter in RMDPs and λ is the robustness penalization parameter in RRMDPs, which are inversely related (Yang et al., 2023, Theorem 3.1). $c(\lambda)$ is some function on λ that varies according to different φ-divergences. N is the dataset size used by algorithms. † The bound of HyTQ is not directly comparable with others in terms of $V_{\max}$ since the non-stationary finite-horizon setting requires H multiplicity in dataset size. $d_{\min}$ is the minimal positive value of data generating stationary distribution d, i.e. $\min_{s,a} d(s,a)$. $\mathcal{F}$ and $\mathcal{G}$ are two function representations, and $(\mathcal{S},\mathcal{A})$ is the state-action space.

a known stochastic reward function, $P \in \Delta(\mathcal{S})^{|\mathcal{S}||\mathcal{A}|}$ is a probability transition function describing an environment, γ is a discount factor, and d_0 is the starting state distribution. A stationary (stochastic) policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ specifies a distribution over actions in each state. We denote the transition dynamic distribution at state-action (s,a) as $P_{s,a} \in \Delta(\mathcal{S})$. For convenience, we write $r(s,a) = \mathbb{E}_{r\sim R(s,a)}[r]$ and assume it is deterministic as in RL literature (Agarwal et al., 2019) since the performance guarantee will be identical up to a constant factor.

The value function of a policy π is $V_{P,r}^{\pi}(s) = \mathbb{E}_{P,\pi}[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t) \mid s_0 = s]$ starting at state $s_0 = s$ and $a_t \sim \pi(s_t), s_{t+1} \sim P_{s_t,a_t}$ for all $t \geq 0$. Similarly, we define an action-value function of a policy π as $Q_{P,r}^{\pi}(s,a) = \mathbb{E}_{P,\pi}[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t) \mid s_0 = s, a_0 = a]$. Each policy π induces a discounted occupancy density over state-action pairs $d_P^{\pi} : \mathcal{S}\times\mathcal{A} \to [0,1]$ defined as $d_P^{\pi}(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P_t(s_t = s, a_t = a;\pi)$, where $P_t(s_t = s, a_t = a;\pi)$ denotes the visitation probability of state-action pair (s,a) at time step t, starting at $s_0 \sim d_0(\cdot)$ and following π on the model P. The optimal policy π_P^* achieves the maximum value of any policy $V_{P,r}^{\pi}$.

Offline Reinforcement Learning: The goal of offline RL on γMDP (P^o, r) is to learn a *good* policy $\hat{\pi}$ (a policy with a high $V_{P^o,r}^{\hat{\pi}}$) based only on the *offline dataset*. An offline dataset is a historical and fixed dataset of interactions $\mathcal{D}_{P^o} = \{(s_i,a_i,s_i')\}_{i=1}^N$, where $s_i' \sim P_{s_i,a_i}^o$ and the (s_i,a_i) pairs are independently and identically generated according to a data distribution $\mu \in \Delta(\mathcal{S}\times\mathcal{A})$. For convenience, μ also denotes the *offline/behavior policy* that generates $\mathcal{D}_{P^o}$. One classical offline RL algorithm with general function approximation capabilities with provable performance guarantees is *Fitted Q-Iteration (FQI)* (Szepesvári & Munos, 2005; Chen & Jiang, 2019; Liu et al., 2020). A function class $\mathcal{F} = \{f : \mathcal{S}\times\mathcal{A} \to [0, 1/(1-\gamma)]\}$ (e.g., neural networks, kernel functions, linear functions, etc) represents Q-value functions of γMDP (P^o, r). At each iteration, given $f_k \in \mathcal{F}$ and $\mathcal{D}_{P^o}$, FQI does the following least-square regression for the approximate squared Bellman error: $f_{k+1} = \arg\min_{f\in\mathcal{F}} \mathbb{E}_{\mathcal{D}_{P^o}}[(y_{f_k} - f)^2]$, where $y_{f_k}(s,a,s') = r(s,a) + \gamma\max_b f_k(s',b)$. In this regression step, FQI aims to find the optimal action-value $Q_{P^o,r}^{\pi^*}$ by approximating the non-robust squared Bellman error ($\|r + \gamma\mathbb{E}_{P^o}V_{P^o,r}^{\pi^*}(\cdot) - Q_{P^o,r}^{\pi^*}\|_{2,\mu}^2$) using offline data $\mathcal{D}_{P^o}$ with function approximation $\mathcal{F}$. Finally, for some starting state $s_0 \sim d_0$, the performance guarantee of an algorithm policy $\hat{\pi}$ is given by bounding the *suboptimality* quantity $0 \leq V_{P^o,r}^{\pi^*}(s_0) - V_{P^o,r}^{\hat{\pi}}(s_0)$.

Infinite-Horizon Robust φ-Regularized Markov Decision Process: Let P^o be the nominal model, that is, a probability transition function describing a training environment. An infinite-horizon discounted Robust φ-Regularized Markov Decision Process (γRRMDP) tuple $(\mathcal{S},\mathcal{A},r,P^o,\lambda,\gamma,\varphi,d_0)$ where $\lambda > 0$ is a robustness parameter and $\varphi : \mathbb{R} \to \mathbb{R}$ is a convex function. The *robust regularized reward function* is defined as $r_P^{\lambda}(s,a) =$

$r(s,a) + \lambda\gamma D_\varphi(P_{s,a}, P^o_{s,a})$ for any state-action pairs and any P such that $P_{s,a}, P^o_{s,a}$. Here D_φ is the φ-divergence (Csiszár, 1967) defined as

$$D_\varphi(p,q) = \int \varphi\left(\mathrm{d}p/\mathrm{d}q\right)\mathrm{d}q \tag{1}$$

for two probability distributions p and q with $p \ll q$, where φ is convex on $\mathbb{R}$ and differentiable on $\mathbb{R}_+$ satisfying $\varphi(1) = 0$ and $\varphi(t) = +\infty$ for $t < 0$. Examples of φ-divergence include Total Variation (TV), Kullback-Leibler (KL), chi-square, Conditional Value at Risk (CVaR), and more (c.f. Proposition 3). The *robust regularized value function* of a policy π is defined as

$$V^\pi_\lambda = \inf_{P\in\mathcal{P}} V^\pi_{P,r^\lambda_P}, \tag{2}$$

where $\mathcal{P} = \otimes_{s,a}\mathcal{P}_{s,a}$ and $\mathcal{P}_{s,a} = \{P_{s,a} \in \Delta(\mathcal{S}) : P_{s,a} \ll P^o_{s,a}, \forall (s,a) \in \mathcal{S}\times\mathcal{A}\}$. By definition, for any π, it follows that $V^\pi_\lambda \le V^\pi_{P^o,r} \le 1/(1-\gamma)$. The *optimal robust regularized value function* is $V^*_\lambda = \max_\pi V^\pi_\lambda$ (similarly we can design Q^*_λ), and π^* is the *robust regularized optimal policy* that achieves this optimal value. For convenience, we denote $V^*_\lambda(Q^*_\lambda)$ as $V^*(Q^*)$. We note that $\mathcal{P}$ satisfies the (s,a)-rectangularity condition (Iyengar, 2005) by definition. This is a sufficient condition for the optimization problem in (2) to be tractable. It also enables the existence of a *deterministic policy* for π^* (Yang et al., 2023). We formally mention this in Proposition 5. For any policy π, denote $V^\pi = \mathbb{E}_{s\sim d_0}[V^\pi(s)]$ as the expected total reward with d_0 as initial state distribution.

Denote the robust regularized Bellman operator $\mathcal{T} : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ as

$$(\mathcal{T}Q)(s,a) = r(s,a) + \gamma \inf_{P_{s,a}\in\mathcal{P}_{s,a}} \left(\mathbb{E}_{s'\sim P_{s,a}}[\max_{a'} Q(s',a')] + \lambda D_\varphi(P_{s,a}, P^o_{s,a})\right). \tag{3}$$

Since $\mathcal{T}$ is a contraction (Yang et al., 2023), the *robust Q-iteration* (RQI) $Q_{k+1} = \mathcal{T}Q_k$ converges to Q^*. We get the robust optimal policy as $\pi^*(s) = \arg\max_a Q^*(s,a)$.

2.1. Problem Conceptualization

In this section, we study the offline infinite-horizon robust φ-regularized RL (γR^{3}L) problem, acquiring useful insights to construct our algorithm (Algorithm 1) in next section. The goal here is to learn a *good* robust policy $\hat{\pi}$ (a policy with a high $V^{\hat{\pi}}_\lambda$) based on the offline dataset. We start by noting one key challenge in the estimation of the robust regularized Bellman operator $\mathcal{T}$ (3): One may require many offline datasets from each $P \in \mathcal{P}$ to achieve our offline γR^{3}L goal. In this work, we use the penalized Distributionally Robust Optimization (DRO) tool (Sinha et al., 2018; Levy et al., 2020; Jin et al., 2021b) to not require such unrealistic existence of offline datasets. In particular, as in non-robust offline RL, we only rely on the offline dataset $\mathcal{D}_{P^o}$ generated on the nominal model P^o by an offline policy μ. This statement is justified via the following proposition.

Proposition 1. *Consider a robust φ-regularized MDP. For any $Q : \mathcal{S}\times\mathcal{A} \to [0, 1/(1-\gamma)]$, the robust regularized Bellman operator $\mathcal{T}$ (3) can be equivalently written as*

$$(\mathcal{T}Q)(s,a) = r(s,a) - \gamma \inf_{\eta\in\Theta}(\lambda\mathbb{E}_{s'\sim P^o_{s,a}}[\varphi^*\left((\eta - V(s'))/\lambda\right)] - \eta), \tag{4}$$

where $V(s) = \max_{a\in\mathcal{A}} Q(s,a)$ and $\Theta \subset \mathbb{R}$ is some bounded real line which depends on φ^*.

A proof of this proposition is given in Appendix E and follows from Levy et al. (2020, Section A.1.2). We refer to (4) as the *robust regularized Bellman dual operator*. Observing the sole dependence on the nominal model P^o in (4), one can come up with estimators for data-driven approaches that naturally depend only on the dataset $\mathcal{D}_{P^o}$. We remark that we consider a class of φ-divergences satisfying the conditions in Proposition 3 for all the results in this paper.

We now remark on a natural first attempt at performing the squared Bellman error least-square regression, like FQI, on the robust regularized Bellman dual operator (4). Observe that the true Bellman error $\mathbb{E}_{s,a\sim\mu}[|\mathcal{T}Q^*(s,a) - Q^*(s,a)|]$ involves solving an inner convex minimization problem in $\mathcal{T}Q^*(s,a)$ (4) for every (s,a). Since we are in a countably large state space regime, it is infeasible to devise approximations to this true squared Bellman error. In addition, we have to also enable general function architecture for action-values. To alleviate this challenging task, we now turn our attention to the inner convex minimization problem in the robust regularized Bellman dual operator (4). Due to the (s,a)-rectangularity assumption, we note that the η's are not correlated across all (s,a). With this note, for every (s,a), we can replace η in $(\mathcal{T}Q)(s,a)$ (4) with a *dual-variable function* $g(s,a)$. Thus, intuitively, multiple point-wise minimizations can be replaced by a single dual-variable functional minimization over the function space of g. We formalize this intuition using *variational functional analysis* (Rockafellar & Wets, 2009) for a countably large state space regime in the following.

We denote $L^1(\mu)$ as the set of all absolutely integrable functions defined on the probability (measure) space $(\mathcal{S}\times\mathcal{A}, \Sigma(\mathcal{S}\times\mathcal{A}), \mu)$ with μ, the data generating distribution, as the σ-finite probability measure. To elucidate, $L^1(\mu)$ is the set of all functions $g : \mathcal{S}\times\mathcal{A} \to \mathcal{C} \subset \mathbb{R}$ such that $\|g\|_{1,\mu}$ is finite. We set $\mathcal{C} = \Theta$ considering the inner minimization in (4). Fixing any given function $f : \mathcal{S}\times\mathcal{A} \to [0, 1/(1-\gamma)]$, we define the loss function $L_{\mathrm{dual}}(g; f)$, for all $g \in L^1(\mu)$, as

$$L_{\mathrm{dual}}(g; f, \mu) = \mathbb{E}_{s,a\sim\mu, s'\sim P^o_{s,a}}[\tag{5}$$

$$\lambda\varphi^*((g(s,a) - \max_{a'} f(s',a'))/\lambda) - g(s,a)].$$

We state the result for single dual-variable functional minimization intuition we developed in the previous paragraph. We also note one variant of this result appears in the distributionally robust RL work (Panaganti et al., 2022).

Proposition 2. *Let L_{dual} be the loss function defined in (5). Then, for any function $f : \mathcal{S} \times \mathcal{A} \to [0, 1/(1-\gamma)]$, we have*

$$\inf_{g \in L^1(\mu)} L_{\text{dual}}(g; f, \mu) = \mathbb{E}_{s,a\sim\mu}\Big[\tag{6}$$
$$\inf_{\eta\in\Theta} (\lambda \mathbb{E}_{s'\sim P^o_{s,a}}[\varphi^*((\eta - \max_{a'} f(s',a'))/\lambda)] - \eta)\Big].$$

We provide a proof in Appendix E, which relies on Rockafellar & Wets (2009, Theorem 14.60).

For any given $f : \mathcal{S}\times\mathcal{A} \to [0, 1/(1-\gamma)]$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$, we define an operator $\mathcal{T}_g$, for all $g \in L^1(\mu)$, as

$$(\mathcal{T}_g f)(s,a) = r(s,a) - \tag{7}$$
$$\gamma(\lambda \mathbb{E}_{s'\sim P^o_{s,a}}[\varphi^*((g(s,a) - V(s'))/\lambda)] - g(s,a)).$$

This operator is useful in view of Propositions 1 and 2. To see this, we first define $g^*(Q) \in \arg\min_{g\in L^1(\mu)} L_{\text{dual}}(g; Q, \mu)$ for any action-value function Q. Now, by taking an expectation w.r.t the data generating distribution μ on (4), we observe $\mathcal{T}Q = \mathcal{T}_{g^*(Q)}Q$ by utilizing (6). Due to this observation, in the following subsection, we develop an algorithm by approximating both the optimal dual-variable function of optimal robust value $g^*(Q^*)$ and the robust squared Bellman error ($\|\mathcal{T}_{g^*(Q^*)}Q^* - Q^*\|^2_{2,\mu}$) using offline data $\mathcal{D}_{P^o}$. Panaganti et al. (2022) similarly conceptualized their total variation φ-divergence robust RL algorithm. Here, Proposition 1 enables us to conceptualize for general φ-divergence.

2.2. Robust φ-regularized fitted Q-iteration

In this section, we formally propose our algorithm based on the tools developed so far. Our proposed algorithm is called Robust φ-regularized fitted Q-iteration (RPQ) Algorithm and is summarized in Algorithm 1. We first discuss the inputs to our algorithm. As mentioned above, we only use the offline dataset $\mathcal{D}_{P^o} = \{(s_i, a_i, s_i')\}_{i=1}^N$, generated according to a data distribution μ on the nominal model P^o. We also consider two general function classes $\mathcal{F} \subset (f : \mathcal{S} \times \mathcal{A} \to [0, 1/(1-\gamma)])$ and $\mathcal{G} \subset (g : \mathcal{S} \times \mathcal{A} \to \Theta)$ representing action-value functions and dual-variable functions, respectively. We now define useful approximation quantities for $g \in \mathcal{G}$ and $f \in \mathcal{F}$. For given f, the empirical loss function of the true loss L_{dual} Equation (5) on $\mathcal{D}_{P^o}$ is

$$\widehat{L}_{\text{dual}}(g; f) = \mathbb{E}_{\mathcal{D}_{P^o}}[\tag{8}$$
$$\lambda\varphi^*((g(s_i,a_i) - \max_{a'} f(s_i',a'))/\lambda) - g(s_i,a_i)].$$

For given f, g, the empirical squared robust regularized Bellman error on $\mathcal{D}_{P^o}$ is

$$\widehat{L}_{\text{robQ}}(Q; f, g) = \mathbb{E}_{\mathcal{D}_{P^o}}[(r(s_i,a_i) - \gamma\lambda\varphi^*((g(s_i,a_i) -$$
$$\max_{a'} f(s_i',a'))/\lambda) + \gamma g(s_i,a_i) - Q(s_i,a_i))^2]. \tag{9}$$

We start with an initial action-value function $Q_0(s,a) = 0$ and execute the following two steps for K iterations. At iteration k of the algorithm with input Q_k, as a first step, we compute a dual-variable function $g_k \in \mathcal{G}$ through the *empirical risk minimization* approach, that is, we solve $\arg\min_{g\in\mathcal{G}} \widehat{L}_{\text{dual}}(g; Q_k)$ *(Line 4 of Algorithm 1)*. As a second step, given inputs Q_k and g_k, we compute the next iterate $Q_{k+1} \in \mathcal{F}$ through the *least-squares regression* method, that is, we solve $\arg\min_{f\in\mathcal{F}} \widehat{L}_{\text{robQ}}(f; Q_k, g_k)$ *(Line 5 of Algorithm 1)*. After K iterations, we extract the greedy policy from Q_K *(Line 7 of Algorithm 1)*.

Algorithm 1 Robust φ-regularized fitted Q-iteration (RPQ) Algorithm

1: **Input:** Regularization φ, offline dataset $\mathcal{D}_{P^o} = (s_i, a_i, r_i, s_i')_{i=1}^N$, general function classes $\mathcal{F}$ and $\mathcal{G}$
2: **Initialize:** $Q_0 \equiv 0 \in \mathcal{F}$.
3: **for** $k = 0, \cdots, K-1$ **do**
4: **Dual variable function minimization:** $g_k = \widehat{g}_{Q_k} = \arg\min_{g\in\mathcal{G}} \widehat{L}_{\text{dual}}(g; Q_k)$ (c.f. (8))
5: **Robust φ-regularized Q-update:** $Q_{k+1} = \arg\min_{Q\in\mathcal{F}} \widehat{L}_{\text{robQ}}(Q; Q_k, g_k)$ (c.f. (9))
6: **end for**
7: **Output:** $\pi_K = \arg\max_a Q_K(s,a)$

2.3. Performance Guarantee: Suboptimality

We now discuss the performance guarantee of our RPQ Algorithm. In particular, we characterize how close the robust regularized value function of our RPQ Algorithm is to the optimal robust regularized value function. We first mention all the assumptions about the data generating distribution μ and the representation power of $\mathcal{F}$ and $\mathcal{G}$ before we present our main results.

Assumption 1 (Concentrability)**.** *There exists a finite constant $C > 0$ such that for any $\nu \in \{d_{\pi,P} \mid$ any policy π and $P \in \mathcal{P}$ satisfying $D_\varphi(P_{s,a}, P^o_{s,a}) \le 1/(\lambda(1-\gamma))$ for all s, a (both can be non-stationary)$\} \subseteq \Delta(\mathcal{S}\times\mathcal{A})$, we have $\|\nu/\mu\|_\infty \le \sqrt{C}$.*

Assumption 1 stipulates the support set of the data generating distribution μ, i.e. $\{(s,a) \in \mathcal{S}\times\mathcal{A} : \mu(s,a) > 0\}$, to cover the union of all support sets of the distributions ν, leading to a *robust exploratory* behavior. This assumption is widely used in the offline RL literature (Munos, 2003; Agarwal et al., 2019; Chen & Jiang, 2019; Wang et al., 2021; Xie et al., 2021) in different forms. We adapt this assumption

from the robust offline RL (Panaganti et al., 2022; Zhang et al., 2024).

Assumption 2 (Approximate Robust Bellman Completeness)**.** *Let $\varepsilon_{\mathcal{F}}$ be some small positive constant. For any $g \in \mathcal{G}$, we have $\sup_{f\in\mathcal{F}} \inf_{f'\in\mathcal{F}} \|f' - \mathcal{T}_g f\|_{2,\mu}^2 \leq \varepsilon_{\mathcal{F}}$ for the data generating distribution μ.*

We note that Assumption 2 holds trivially if $\mathcal{T}_g$ is closed under $\mathcal{F}$, that is, for any $f \in \mathcal{F}$ and $g \in \mathcal{G}$, if it holds that $\mathcal{T}_g f \in \mathcal{F}$, then $\varepsilon_{\mathcal{F}} = 0$. This assumption has been widely used in different forms in the non-robust offline RL literature (Agarwal et al., 2019; Wang et al., 2021; Xie et al., 2021) and robust offline RL literature (Panaganti et al., 2022; Bruns-Smith & Zhou, 2023; Zhang et al., 2024).

Assumption 3 (Approximate Dual Realizability)**.** *For all $f \in \mathcal{F}$, there exists a uniform constant $\varepsilon_{\mathcal{G}}$ such that $\inf_{g\in\mathcal{G}} L_{\text{dual}}(g; f) - \inf_{g\in L^1(\mu)} L_{\text{dual}}(g; f) \leq \varepsilon_{\mathcal{G}}$.*

Assumption 3 holds trivially if $g^*(f) \in \mathcal{G}$ for any $f \in \mathcal{F}$ (since $\varepsilon_{\mathcal{G}} = 0$). This assumption has been used in earlier robust offline RL literature (Panaganti et al., 2022; Bruns-Smith & Zhou, 2023).

Now we state our main theoretical result on the performance of the RPQ algorithm. In Appendix E we restate the result including the constant factors.

Theorem 1. *Let Assumptions 1 to 3 hold. Let $c_\varphi(\lambda, \gamma)$ be problem-dependent constants for φ. Let π_K be the RPQ algorithm policy after K iterations. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have*

$$V^{\pi^*} - V^{\pi_K} \leq \frac{\sqrt{C}(\gamma^K + \sqrt{6\varepsilon_{\mathcal{F}}} + \gamma\varepsilon_{\mathcal{G}})}{(1-\gamma)^2} + \frac{c_\varphi(\lambda, \gamma)}{(1-\gamma)^3}\mathcal{O}(\sqrt{C \log(|\mathcal{F}||\mathcal{G}|/\delta)/N}).$$

Theorem 1 states that the RPQ algorithm is approximately optimal. This theorem also gives the sample complexity guarantee for finding an ε-suboptimal policy w.r.t. the optimal policy π^*. To see this, by neglecting the first term due to inevitable function class approximation errors, for $N \geq \mathcal{O}(\frac{(c_\varphi(\lambda,\gamma))^2}{\varepsilon^2(1-\gamma)^4} \log \frac{|\mathcal{F}||\mathcal{G}|}{\delta})$ we get $V^{\pi^*} - V^{\pi_K} \leq \varepsilon/(1-\gamma)$ with probability at least $1 - \delta$ for any fixed $\varepsilon, \delta \in (0, 1)$.

Remark 1. *Note that the guarantee for the TV case in Theorem 1 requires making another assumption on the existence of a fail-state (Panaganti et al., 2022, Lemma 3), Assumption 8 replacing H with $1/(1-\gamma)$. However, we specialize Theorem 1 for the TV case by relaxing Assumption 1 to get the same guarantee, which we present in Appendix E. In particular, we relax Assumption 1 to the non-robust offline RL concentrability assumption (Foster et al., 2022), i.e. we only need the distribution ν to be in the collection of discounted state-action occupancies on the nominal model P^o.*

3. Hybrid Robust φ-Regularized Reinforcement Learning

In this section, we provide a *hybrid* robust φ-Regularized RL protocol to overcome the out-of-data-distribution issue in offline robust RL. As in Song et al. (2023), we reformulate the problem in the finite-horizon setting to use its backward induction feature that enables RPQ iterates to run in each episode. We again start by discussing preliminaries and the problem formulation.

Finite-Horizon Markov Decision Process: A finite-horizon Markov Decision Process (hMDP) is $(\mathcal{S}, \mathcal{A}, P = (P_h)_{h=0}^{H-1}, r = (r_h)_{h=0}^{H-1}, H)$, where H is the horizon length, for any $h \in [H]$, $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a known deterministic reward function and $P_h \in \Delta(\mathcal{S})^{|\mathcal{S}||\mathcal{A}|}$ is the transition probability function at time h. A non-stationary (stochastic) policy $\pi = (\pi_h)_{h=0}^{H-1}$ where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$. We denote the transition dynamic distribution at time h and state-action (s, a) as $P_{h,s,a} \in \Delta(\mathcal{S})$. Given π, we define the state and action value functions in the usual manner: $V_{P,r}^{h,\pi}(s) = \mathbb{E}[\sum_{t=h}^{H-1} r_t(s_t, a_t)|s_h = s]$ starting at state $s_h = s$ and $a_t \sim \pi_t(s_t), s_{t+1} \sim P_{t+1,s_t,a_t}$, and $Q_{P,r}^{h,\pi}(s, a) = \mathbb{E}[\sum_{t=h}^{H-1} r_t(s_t, a_t)|s_h = s, a_h = a]$ starting at state-action $s_h = s, a_h = a$ and $s_{t+1} \sim P_{t+1,s_t,a_t}, a_{t+1} \sim \pi_{t+1}(s_{t+1})$. Given π, occupancy measure over state-action pairs $d_P^{h,\pi}(s, a) = P_h(s_h = s, a_h = a; \pi)$. We write $\pi_P^* = (\pi_h^*)_{h=0}^{H-1}$ to denote an optimal deterministic policy, which maximizes $V_{P,r}^{\pi} = (V_{P,r}^{h,\pi})_{h=0}^{H-1}$.

Hybrid Reinforcement Learning: The goal of hybrid RL on hMDP (P^o, r) is to learn a good policy $\hat{\pi}$ based on adaptive datasets consisting of *both* offline datasets and on-policy datasets. Given timestep $h \in [H]$, offline dataset $\mathcal{D}_{h,P^o}^{\mu} = \{(s_i, a_i, s_i')_{i=1}^{m_{\text{off}}}\}$ is generated by $s_i' \sim P_{h,s_i,a_i}^o$ with the (s_i, a_i) pairs i.i.d. sampled by $\mu_h \in \Delta(\mathcal{S} \times \mathcal{A})$ offline data distribution. For convenience, $\mu = (\mu_h)_{h=0}^{H-1}$ also denotes the *offline policy* that generates $\mathcal{D}_{P^o}^{\mu}$. Given timestep $h \in [H]$, on-policy dataset $\mathcal{D}_{h,P^o}^{\pi} = \{(s_i, a_i, s_i')_{i=1}^{m_{\text{on}}}\}$ is generated by $(s_i, a_i) \sim d_{P^o}^{h,\pi}$ and $s_i' \sim P_{h,s_i,a_i}^o$ for all the previously learned policies π by the algorithm. Song et al. (2023) proposes *Hybrid Q-learning* (HyQ) algorithm with general function approximation capabilities and provable guarantees for hybrid RL. The HyQ algorithm (c.f. Song et al. (2023, Algorithm 1)) is quite straightforward: For each iteration $k \in [K]$, do backward induction of the FQI algorithm on timesteps $h \in [H]$ using the adaptive datasets described above. Finally, for some starting state $s_0 \sim d_0$, the performance guarantee of algorithm policies $\{\pi_k\}_{k\in[K]}$ is given by bounding the *cumulative suboptimality* quantity $0 \leq \sum_{k=[K]}[V_{P^o,r}^{0,\pi^*}(s_0) - V_{P^o,r}^{0,\pi_k}(s_0)]$. We note the total adaptive dataset size is N to provide comparable results with offline RL.

Finite-Horizon Robust φ-Regularized Markov Decision

Process: Again, let P^o be the nominal model. A finite-horizon discounted Robust φ-Regularized Markov Decision Process (hRRMDP) tuple $(\mathcal{S}, \mathcal{A}, P^o = (P_h^o)_{h=0}^{H-1}, r = (r_h)_{h=0}^{H-1}, \lambda, H, \varphi, d_0)$ where $\lambda > 0$ is a robustness parameter and $\varphi : \mathbb{R} \to \mathbb{R}$ is as before. For $h \in [H]$, the *robust regularized reward function* is $r_h^\lambda(s,a) = r_h(s,a) + \lambda D_\varphi(P_{h,s,a}, P^o_{h,s,a})$. For $h \in [H]$, the *robust regularized value function* of a policy π is defined as $V^\pi_{h,\lambda} = \inf_{P\in\mathcal{P}} V^{h,\pi}_{P,r_h^\lambda}$, where $\mathcal{P} = \otimes_{h,s,a}\mathcal{P}_{h,s,a}$ and $\mathcal{P}_{h,s,a} = \{P_{h,s,a} \in \Delta(\mathcal{S}) : P_{h,s,a} \ll P^o_{h,s,a}, \forall (s,a) \in \mathcal{S}\times\mathcal{A}$ and $h \in [H]\}$. By definition, for any π, it follows that $V^\pi_{h,\lambda} \le V^{h,\pi}_{P^o,r} \le H$. For $h \in [H]$, the *optimal robust regularized value function* is $V^*_{h,\lambda} = \max_\pi V^\pi_{h,\lambda}$, and π^* is the *robust regularized optimal policy* that achieves this optimal value. For convenience, we denote $V^*_{h,\lambda}(Q^*_{h,\lambda})$ as $V^*_h(Q^*_h)$ for all $h \in [H]$. We again note that, for each $h \in [H]$, $\mathcal{P}$ satisfies the (s,a)-rectangularity condition (Iyengar, 2005) by definition. It enables the existence of a *non-stationary deterministic policy* for π^* (Zhang et al., 2024). We formalize this in Proposition 6. We denote $V^\pi = \mathbb{E}_{s\sim d_0}[V_0^\pi(s)]$ as the expected total reward.

For convenience, we let $Q^\pi_{H,\lambda} = 0$ for any π. For any $h \in [H]$, denote the robust regularized Bellman operator $\mathcal{T} : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ as

$$(\mathcal{T}Q_{h+1})(s,a) = r_h(s,a) + \inf_{P_{h,s,a}\in\mathcal{P}_{h,s,a}} \big(\mathbb{E}_{s'\sim P_{h,s,a}}[\max_{a'} Q_{h+1}(s',a')] + \lambda D_\varphi(P_{h,s,a}, P^o_{h,s,a})\big). \tag{10}$$

As $Q^*_H = 0$, doing backward iteration of $\mathcal{T}$, i.e., the *robust dynamic programming* $Q^*_h = \mathcal{T}Q^*_{h+1}$, we get Q^*_h for all $h \in [H]$. For each timestep $h \in [H]$, we also get the robust optimal policy as $\pi^*_h(s) = \arg\max_a Q^*_h(s,a)$.

3.1. Problem Conceptualization

In this section, we study the hybrid finite-horizon robust TV-regularized RL problem, acquiring the necessary insights to construct our algorithm (Algorithm 2) in the next section. We conceptualize for general φ-divergence, but only propose our algorithm for total variation φ-divergence. The goal here is to learn a *good* robust policy $\hat{\pi}$ based on adaptive datasets consisting of *both* offline datasets and on-policy datasets. We start by noting a direct consequence of Proposition 1 due to similar inner minimization problems in both infinite horizon (3) and finite horizon (10) operators.

Corollary 1. *For any $Q_h : \mathcal{S}\times\mathcal{A} \to [0,H]$ and $h \in [H]$, the robust regularized Bellman operator $\mathcal{T}$ (10) can be equivalently written as*

$$(\mathcal{T}Q_{h+1})(s,a) = r_h(s,a) - \gamma \inf_{\eta\in\Theta} (\lambda \mathbb{E}_{s'\sim P^o_{h,s,a}}[\varphi^*((\eta - V_{h+1}(s'))/\lambda)] - \eta), \tag{11}$$

where $V_{h+1}(s) = \max_{a\in\mathcal{A}} Q_{h+1}(s,a)$ and $\Theta \subset \mathbb{R}$ is some bounded real line that depends on φ^*.

As in Section 2, this dual reformulation enables us to use the datasets from only the nominal model P^o for estimating the robust regularized operator in its primal form (10).

We start by recalling the philosophy of the HyQ algorithm (Song et al., 2023) to use the FQI algorithm for adaptive datasets. We do the same for our hybrid finite-horizon robust φ-regularized RL problem here. For each $h \in [H]$, we need to estimate the true Bellman error $\mathbb{E}_{s,a\sim\mu_h}[|\mathcal{T}Q^*_{h+1}(s,a) - Q^*_h(s,a)|] + \sum_{t=0}^{k-1} \mathbb{E}_{s,a\sim d^{\pi_t}_{h,P^o}}[|\mathcal{T}Q^*_{h+1}(s,a) - Q^*_h(s,a)|]$ using offline dataset from μ_h and the on-policy dataset from $d^{\pi_t}_{h,P^o}$ by the learned policies from the algorithm. We remark that the out-of-data-distribution issue appears when we only have access to the offline dataset to estimate the summation term above, which depends on $d^{\pi_t}_{h,P^o}$.

As discussed in Section 2, the true Bellman error itself involves solving an inner convex minimization problem in $\mathcal{T}Q^*_{h+1}(s,a)$ (11) for every (s,a) and h that is challenging for countably large state setting. To alleviate this challenging task, we again utilize the functional minimization Proposition 2 developed in Section 2. For any h, we denote the set of *admissible distributions* of nominal model P^o as $\mathbb{D}_h = \{\mu_h\} \cup \{d^\pi_{h,P^o} \,|\, \text{for any policy (including non-stationary) } \pi\}$. Now we redefine dual loss for any $f_{h+1} \in \mathcal{F}_{h+1}, \nu_h \in \mathbb{D}_h$, as

$$L_{\text{dual}}(g; f_{h+1}, \nu_h) = \mathbb{E}_{s,a\sim\nu_h, s'\sim P^o_{h,s,a}}[\lambda\varphi^*((g(s,a) - \max_{a'} f_{h+1}(s',a'))/\lambda) - g(s,a)]. \tag{12}$$

We state a direct consequence of Proposition 2 here.

Corollary 2. *Let L_{dual} be the loss function defined in (12). Fix $h \in [H]$ and consider any policy π. Then, for any function $f_{h+1} : \mathcal{S}\times\mathcal{A} \to [0,H]$ and any $\nu_h \in \mathbb{D}_h$, we have*

$$\inf_{g\in L^1(\nu_h)} L_{\text{dual}}(g; f_{h+1}, \nu_h) = \mathbb{E}_{s,a\sim\nu_h}\Big[\inf_{\eta\in\Theta}(\lambda\mathbb{E}_{s'\sim P^o_{h,s,a}}[\varphi^*((\eta - \max_{a'} f_{h+1}(s',a'))/\lambda)] - \eta)\Big]. \tag{13}$$

For any given $f_h : \mathcal{S}\times\mathcal{A} \to [0,H]$ and h, we redefine operator $\mathcal{T}_g$ for all $g \in \mathcal{G}_h$, as

$$(\mathcal{T}_g f_{h+1})(s,a) = r_h(s,a) - \lambda\mathbb{E}_{s'\sim P^o_{h,s,a}}[\varphi^*((g(s,a) - \max_{a'} f_{h+1}(s',a'))/\lambda)] + g(s,a). \tag{14}$$

We have all the necessary tools now. In the following subsection, we develop an algorithm that naturally extends our RPQ algorithm using adaptive datasets.

3.2. Hybrid Robust regularized Q-iteration

In this section, we propose our algorithm based on the tools developed so far. Our proposed algorithm is called Hybrid robust Total-variation-regularized Q-iteration (HyTQ:

Algorithm 2 HyTQ Algorithm

Input: Offline dataset $\mathcal{D}_h^\mu \sim \mu_h$ of size $m_{\text{off}} = T$ for $h \in [H]$, general function classes $\mathcal{F}$ and $\mathcal{G}$.
Initialize: $Q_h^0 \equiv 0 \in \mathcal{F}_h$.
for $k = 0, \cdots, K-1$ **do**
  Compute π_k as $\pi_{k,h}(s) = \arg\max_a Q_h^k(s,a)$
  $\forall h$, collect m_{on}=1 online dataset $\mathcal{D}_h^k \sim d_{h,P^o}^{\pi_k}$
  Initialize: $Q_H^{k+1} \equiv 0 \in \mathcal{F}_H$
  for $h = H-1, \cdots, 0$ **do**
    Aggregate adaptive dataset $\mathcal{D}_h^k = \mathcal{D}_h^\mu + \sum_{\tau=0}^k \mathcal{D}_h^\tau$
    Dual variable function minimization: (c.f. (15)) $g_h^{k+1} = \arg\min_{g\in\mathcal{G}_h} \widehat{L}_{\text{dual}}(g; Q_{h+1}^{k+1}, \mathcal{D}_h^k)$
    Robust φ-regularized Q-update: (c.f. (16)) $Q_h^{k+1} = \arg\min_{Q\in\mathcal{F}_h} \widehat{L}_{\text{robQ}}(Q; Q_{h+1}^{k+1}, g_h^{k+1}, \mathcal{D}_h^k)$
  end for
end for

pronounced *height-Q*) Algorithm, summarized in Algorithm 2. The total variation D_{TV} φ-divergence (1) is defined with $\varphi(t) = |t-1|/2$. The inputs to this algorithm are the offline dataset, and two general function classes $\mathcal{F} = \otimes_{h\in[H]}\mathcal{F}_h, \mathcal{G} = \otimes_{h\in[H]}\mathcal{G}_h$. For any $h \in [H]$, $\mathcal{F}_h \subset (f : \mathcal{S}\times\mathcal{A} \to [0,H])$ and $\mathcal{G}_h \subset (g : \mathcal{S}\times\mathcal{A} \to [0,\lambda])$ represent action-value functions and dual-variable functions at h, respectively. We redefine, using (17), the empirical dual loss and the robust empirical squared robust regularized Bellman error for dataset $\mathcal{D}$ as

$$\widehat{L}_{\text{dual}}(g; f, \mathcal{D}) = \mathbb{E}_{\mathcal{D}}[(g(s_i,a_i) - \max_{a'} f(s_i', a'))_+ - g(s_i,a_i)] \quad \text{and} \tag{15}$$

$$\widehat{L}_{\text{robQ}}(Q; f, g, \mathcal{D}) = \mathbb{E}_{\mathcal{D}}[(r_h(s_i,a_i) - (g(s_i,a_i) - \max_{a'} f(s_i',a'))_+ + g(s_i,a_i) - Q(s_i,a_i))^2]. \tag{16}$$

3.3. Cumulative Suboptimality Guarantee

We now discuss the performance guarantee in terms of the cumulative suboptimality of our HyTQ Algorithm. We first mention all the assumptions before we present our main result and add a brief discussion. We provide detailed discussion in Section 4.

Assumption 4 (Robust Bellman Error Transfer Coefficient)**.** *Let $\mu_h \in \Delta(\mathcal{S}\times\mathcal{A})$ be the offline data generating distribution. For any $f \in \mathcal{F}$, there exists a small positive constant $C(\pi^*)$ for the optimal policy π^* that satisfies*

$$\frac{\sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_{P^o}^{h,\pi^*}}[\mathcal{T} f_{h+1}(s,a) - f_h(s,a)]}{\sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim\mu_h}[|\mathcal{T} f_{h+1}(s,a) - f_h(s,a)|]} \le C(\pi^*).$$

We develop this assumption from non-robust offline RL work (Song et al., 2023).

Assumption 5 (Approximate Value Realizability and Robust Bellman Completeness)**.** *Let $\varepsilon_{\mathcal{F},\text{r}} \ge 0$ be small constant. For any $h \in [H]$ and $g_h \in \mathcal{G}_h$, we have $\inf_{f\in\mathcal{F}_h} \sup_{\nu_h} \|f - \mathcal{T}_{g_h} f_{h+1}\|_{2,\nu_h}^2 \le \varepsilon_{\mathcal{F},r}$ for all $\nu_h \in \mathbb{D}_h$. Furthermore, for any $f_{h+1} \in \mathcal{F}_{h+1}$, we have $\mathcal{T}_{g_h} f_{h+1} \in \mathcal{F}_h$.*

Assumption 6 (Approximate Dual Realizability)**.** *Let $\varepsilon_{\mathcal{G}}$ be some small positive constant. For any $h \in [H]$ and $f_{h+1} \in \mathcal{F}_{h+1}$, we have $\inf_{g\in\mathcal{G}_h} L_{\text{dual}}(g; f_{h+1}, \nu_h) - \inf_{g\in L^1(\nu_h)} L_{\text{dual}}(g; f_{h+1}, \nu_h) \le \varepsilon_{\mathcal{G}}$, for all $\nu_h \in \mathbb{D}_h$.*

We adapt these two enhanced realizability assumptions from the non-robust offline RL literature (Xie et al., 2021; Foster et al., 2022; Song et al., 2023) to our problem. The assumptions in Section 2 are not directly comparable, but for the sake of exposition, let $\mathcal{F}_h, \mathcal{G}_h$ be the same across h. First, note that Assumption 3 with all-policy concentrability (Assumption 1) is equivalent to Assumption 6. Second, Assumption 2 implies $\inf_{f\in\mathcal{F}} \|f - \mathcal{T}_g f\|_{2,\mu}^2 \le \varepsilon_{\mathcal{F}}$. Now again, with all-policy concentrability (Assumption 1), it is the approximate value realizability (Assumption 5). We know non-robust offline RL is hard (Foster et al., 2022) with just realizability and all-policy concentrability. As robust RL is at least as hard as its non-robust counterpart (Panaganti & Kalathil, 2022), we also assume Bellman completeness in Assumption 5.

Assumption 7 (Bilinear Models)**.** *Consider any $f \in \mathcal{F}, g \in \mathcal{G}$ and $h \in [H]$. Let π^f be greedy policy w.r.t f. There exists an unknown feature mapping $X_h : \mathcal{F} \mapsto \mathbb{R}^d$ and two unknown weight mappings $W_h^{\text{q}}, W_h^{\text{d}} : \mathcal{F}\times\mathcal{G} \mapsto \mathbb{R}^d$ with $\max_f \|X_h(f)\|_2 \le B_X$ and $\max_{f,g} \max\{\|W_h^{\text{q}}(f,g)\|_2, \|W_h^{\text{d}}(f,g)\|_2\} \le B_W$ such that both $\mathbb{E}_{d_h^{\pi^f}}[(f_h(s,a) - T_{g_h} f_{h+1})_+] = |\langle X_h(f), W_h^{\text{q}}(f,g)\rangle|$ and $\mathbb{E}_{d_h^{\pi^f}}[(T_{g_h} f_{h+1} - T f_{h+1})_+] = \left|\langle X_h(f), W_h^{\text{d}}(f,g)\rangle\right|$ holds.*

We adapt this problem architecture assumption on P^o with $\mathcal{F}$ and $\mathcal{G}$ for our setting from a series of non-robust online RL works (Jin et al., 2021a; Du et al., 2021).

Assumption 8 (Fail-state)**.** *There is a fail state $s_{f,h}$ for all $h \in [H]$, such that $r_h(s_f, a) = 0$ and $P_{h,s_f,a}(s_{f,h}) = 1$, for all $a \in \mathcal{A}$ and $P \in \mathcal{P}$ satisfying $D_{\text{TV}}(P_{h',s',a'}, P_{h',s',a'}^o) \le \max\{1, H/\lambda\}$ for all h', s', a'.*

This assumption enables us to ground the value of such P's at $s_{f,h}$ to zero, which helps us to get a tight duality (c.f. (17)) without having to know the minimum value across large $\mathcal{S}$. There are approximations to this in the literature (Wang & Zou, 2022). But we adopt this less restrictive assumption from Panaganti et al. (2022) for convenience.

Now we state our main theoretical result on the performance of the HyTQ algorithm. The proof is presented in Appendix F.

Theorem 2. *Let Assumptions 4 to 8 hold. Fix any $\delta \in (0,1)$. Then, HyTQ algorithm policies $\{\pi_k\}_{k\in[K]}$ satisfy $\sum_{k=0}^{K-1}(V^{\pi^*} - V^{\pi_k}) \le \widetilde{\mathcal{O}}(\sqrt{\varepsilon_{\mathcal{F},\text{r}}} + \varepsilon_{\mathcal{G}}) +$*

$\widetilde{\mathcal{O}}(\max\{C(\pi^*),1\}\sqrt{dH^2K}(\lambda+H)\log(|\mathcal{F}||\mathcal{G}|/\delta))$ *with probability at least* $1-\delta$.

Remark 2. *We specialize this result for bilinear model examples,* linear occupancy complexity model *(Du et al., 2021, Definition 4.7) and* low-rank feature selection model *(Du et al., 2021, Definition A.1), in Appendix F.2. We also specialize this result using standard online-to-batch conversion (Shalev-Shwartz & Ben-David, 2014) for uniform policy over HyTQ policies* $\{\pi_k\}_{k\in[K]}$ *to provide sample complexity* $\widetilde{\mathcal{O}}(\max\{(C(\pi^*))^2,1\}dH^3(\lambda+H)^2(\log(|\mathcal{F}||\mathcal{G}|/\delta))^2)/\varepsilon^2$ *in the Appendix F.2.*

4. Theoretical Discussions and Final Remarks

In this section, we compare our results with the most relevant ones from the robust RL literature for the total variation φ-divergence setting. Our Table 1 should be used as a reference. We provide more detailed discussions in Appendix B on the proof ideas of our results, comparison of results for other φ-divergence specializations, and the bilinear model architecture used in the hybrid robust RL setting.

As mentioned in Remark 1, we have a specialized result in Appendix E.2 for the total variation φ-divergence. We get the suboptimality result (Theorem 4) for the RPQ algorithm as $\widetilde{\mathcal{O}}\left(\frac{\lambda\sqrt{C_{\text{tv}}\log(|\mathcal{F}||\mathcal{G}|)}}{(1-\gamma)^3\sqrt{N}}\right)$, where we only have presented the higher-order terms. Panaganti et al. (2022, Theorem 1) mentioned in Table 1 also exhibits same suboptimality guarantee replacing λ with ρ^{-1}. As we noted before, ρ (the robustness radius parameter in RMDPs) and λ (the robustness penalization parameter in RRMDPs) are inversely related (Yang et al., 2023), and for the TV φ-divergence we observe a straightforward relation between the two as $\lambda=\rho^{-1}$. Now for a tabular setting bound, our result further reduces to $\widetilde{\mathcal{O}}\left(\frac{\lambda|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\sqrt{N}}\right)$. Now comparing this to the minimax lower bound (Shi et al., 2023, Theorem 2), our suboptimality bound is worse off by the factors $\sqrt{|\mathcal{S}||\mathcal{A}|}$ and $1/(1-\gamma)$. Nevertheless, we push the boundaries by providing novel suboptimality guarantee studying the robust RL problem in the hybrid RL setting. Furthermore, as mentioned earlier in Remark 2, we provide the offline+online robust RL suboptimality guarantee $\widetilde{\mathcal{O}}\left(\max\{C(\pi^*),1\}\sqrt{dH^3}(\lambda+H)\log(|\mathcal{F}||\mathcal{G}|/\delta)/\sqrt{N}\right)$ in the Appendix F. We also remark that the HyTQ algorithm can be proposed under the RMDP setting with a similar suboptimality guarantee due to the similarity of the dual Bellman equations under the TV φ-divergence for RMDPs and RRMDPs (c.f. Equation (33) and Xu* et al. (2023, Lemma 8)). For the sake of consistency and novelty, we present our results solely for the RRMDP setting. As mentioned earlier, the concentrability assumption improvement is two-fold (Lemma 8): all-policy concentrability (Assumption 9) to single concentrability to transfer coefficient. This is the first of its kind result that does not yet have any existing lower bounds to compare in the robust RL setting. Under similar transfer coefficient, Bellman completeness, and bilinear model assumptions, the HyTQ algorithm sample complexity (Corollary 5) is comparable to that of a non-robust RL algorithm (Song et al., 2023), i.e., $\widetilde{\mathcal{O}}(\max\{(C(\pi^*))^2,1\}dH^5\log(H|\mathcal{F}|/\delta)/\varepsilon^2)$. We leave it to future work for developing minimax rates and getting optimal algorithm guarantees.

We also offer computational tractability in our RPQ and HyTQ algorithms due to the usage of empirical risk minimization (Steps 4 & 9 resp.), over the general function class $\mathcal{G}$, and least-squares (Steps 5 & 10 resp.), over the general function class $\mathcal{F}$, *computationally tractable* estimators. This two-step estimator update avoids the complexity of solving the inner problem for each state-action pair (leading to scaling issues for high-dimensional problems) in the original robust Bellman operators (Equations (3) and (10)). We conclude this section with an exciting future research direction that remains unsolved in this paper. To solve the hybrid robust RL problem for general φ-divergence. In this work, we noticed while building hybrid learning for robust RL that one would require online samples from the worse-case model (c.f. the model that solves the inner problem in robust Bellman operator Equation (10)) for general φ-divergences due to the current analyses dependent on the bilinear models. We use the dual reformulation for the total variation φ-divergence and provide current results supporting the HyTQ algorithm. We remark that using the same approach for other general φ-divergences, we get exponential dependence on the horizon factor. This warrants more sophisticated algorithm designs for the hybrid robust RL problem under general φ-divergences.

5. Conclusion

In this work, we presented two robust RL algorithms. We proposed Robust φ-divergence-fitted Q-iteration algorithm for general φ-divergence in the offline RL setting. We provided performance guarantees with unified analysis for all φ-divergences with arbitrarily large state space using function approximation. To mitigate the out-of-data-distribution issue by improving the assumptions on data generation, we proposed a novel framework called hybrid robust RL that uses both offline and online interactions. We proposed the Total-variation-divergence Q-iteration algorithm in this framework with an accompanying guarantee. We have provided our theoretical guarantees in terms of suboptimality and sample complexity for both offline and offline+online robust RL settings. We also rigorously specialized our results to different φ-divergences and different bilinear modeling assumptions. We have provided detailed comparisons with relevant prior works while also discussing interesting future directions in the field of robust reinforcement learning.

Acknowledgment

KP acknowledges support from the 'PIMCO Postdoctoral Fellow in Data Science' fellowship at the California Institute of Technology. This work acknowledges support from NSF CNS-2146814, CPS-2136197, CNS-2106403, NGSDI-2105648, and funding from the Resnick Institute. EM acknowledges support from NSF award 2240110. We thank several anonymous ICML 2024 reviewers for their constructive comments on an earlier draft of this paper.

Impact Statement

This paper presents work that aims to advance the field of Robust Reinforcement Learning for learning robust policies against model parameter mismatches. This work is of a rigorous theoretical nature; hence, the potential societal consequences of our work do not exist, or none of which we feel must be specifically highlighted here.

References

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 2019.

Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

Bertsimas, D., Gupta, V., and Kallus, N. Data-driven robust optimization. *Math. Program.*, 167(2): 235–292, feb 2018. ISSN 0025-5610. doi: 10.1007/s10107-017-1125-8. URL https://doi.org/10.1007/s10107-017-1125-8.

Blanchet, J., Kang, Y., and Murthy, K. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019. doi: 10.1017/jpr.2019.49.

Blanchet, J., Lu, M., Zhang, T., and Zhong, H. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *Advances in Neural Information Processing Systems*, 36, 2023.

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., and Hassabis, D. Reinforcement learning, fast and slow. *Trends in cognitive sciences*, 23(5):408–422, 2019.

Bruns-Smith, D. and Zhou, A. Robust fitted-q-evaluation and iteration under sequentially exogenous unobserved confounders. *arXiv preprint arXiv:2302.00662*, 2023.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051, 2019.

Chen, J., Patton, R. J., and Zhang, H.-Y. Design of unknown input observers and robust fault detection filters. *International Journal of control*, 63(1):85–105, 1996.

Chen, R., Paschalidis, I. C., et al. Distributionally robust learning. *Foundations and Trends® in Optimization*, 4 (1-2):1–243, 2020.

Chen, Z., Khodadadian, S., and Maguluri, S. T. Finite-sample analysis of off-policy natural actor–critic with linear function approximation. *IEEE Control Systems Letters*, 6:2611–2616, 2022.

Corporation, N. Closing the sim2real gap with nvidia isaac sim and nvidia isaac replicator, 2021. URL https://rb.gy/6xcwgi/.

Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

Du, S., Kakade, S., Lee, J., Lovett, S., Mahajan, G., Sun, W., and Wang, R. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pp. 2826–2836, 2021.

Duchi, J. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

Farahmand, A.-m., Szepesvári, C., and Munos, R. Error propagation for approximate policy and value iteration. *Advances in Neural Information Processing Systems*, 23, 2010.

Fawzi, A., Balog, M., Huang, A., Hubert, T., Romera-Paredes, B., Barekatain, M., Novikov, A., R Ruiz, F. J., Schrittwieser, J., Swirszcz, G., et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53, 2022.

Foster, D. J., Krishnamurthy, A., Simchi-Levi, D., and Xu, Y. Offline reinforcement learning: Fundamental barriers for value function approximation. *COLT, arXiv preprint arXiv:2111.10919*, 2022.

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

Fujimoto, S., Meger, D., and Precup, D. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062, 2019.

Gao, R. and Kleywegt, A. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 2022.

Huang, A., Chen, J., and Jiang, N. Reinforcement learning in low-rank mdps with density features. In *International Conference on Machine Learning*, pp. 13710–13752, 2023.

Iyengar, G. N. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

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

Jin, J., Zhang, B., Wang, H., and Wang, L. Non-convex distributionally robust optimization: Non-asymptotic analysis. *Advances in Neural Information Processing Systems*, 34:2771–2782, 2021b.

Kostrikov, I., Fergus, R., Tompson, J., and Nachum, O. Offline reinforcement learning with fisher divergence critic regularization. In *International Conference on Machine Learning*, pp. 5774–5783, 2021.

Kumar, A., Fu, J., Soh, M., Tucker, G., and Levine, S. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, pp. 11784–11794, 2019.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 1179–1191, 2020.

Lange, S., Gabel, T., and Riedmiller, M. Batch reinforcement learning. In *Reinforcement learning*, pp. 45–73. Springer, 2012.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Lesort, T., Lomonaco, V., Stoian, A., Maltoni, D., Filliat, D., and Díaz-Rodríguez, N. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Information fusion*, 58:52–68, 2020.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33: 8847–8860, 2020.

Liang, Z., Ma, X., Blanchet, J., Zhang, J., and Zhou, Z. Single-trajectory distributionally robust reinforcement learning. *arXiv preprint arXiv:2301.11721*, 2023.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Provably good batch off-policy reinforcement learning without great exploration. In *Neural Information Processing Systems*, 2020.

Liu, Z., Bai, Q., Blanchet, J., Dong, P., Xu, W., Zhou, Z., and Zhou, Z. Distributionally robust q-learning. In *International Conference on Machine Learning*, pp. 13623–13643, 2022.

Mankowitz, D. J., Levine, N., Jeong, R., Abdolmaleki, A., Springenberg, J. T., Shi, Y., Kay, J., Hester, T., Mann, T., and Riedmiller, M. Robust reinforcement learning for continuous control with model misspecification. In *International Conference on Learning Representations*, 2020.

Mannor, S., Mebel, O., and Xu, H. Robust mdps with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509, 2016.

Maraun, D. Bias correcting climate change simulations-a critical review. *Current Climate Change Reports*, 2: 211–220, 2016.

Mirhoseini, A., Goldie, A., Yazgan, M., Jiang, J. W., Songhori, E., Wang, S., Lee, Y.-J., Johnson, E., Pathak, O., Nazi, A., et al. A graph placement methodology for fast chip design. *Nature*, 594(7862):207–212, 2021.

Munos, R. Error bounds for approximate policy iteration. In *ICML*, volume 3, pp. 560–567, 2003.

Munos, R. Performance bounds in l_p-norm for approximate value iteration. *SIAM journal on control and optimization*, 46(2):541–561, 2007.

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

Namkoong, H. and Duchi, J. C. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29, 2016.

Nilim, A. and El Ghaoui, L. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Panaganti, K. *Robust Reinforcement Learning: Theory and Algorithms*. PhD thesis, Texas A&M University, 2023.

Panaganti, K. and Kalathil, D. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *International Conference on Machine Learning (ICML)*, pp. 511–520, 2021a.

Panaganti, K. and Kalathil, D. Sample complexity of model-based robust reinforcement learning. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pp. 2240–2245, 2021b.

Panaganti, K. and Kalathil, D. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 9582–9602, 2022. URL https://proceedings.mlr.press/v151/panaganti22a.html.

Panaganti, K., Xu, Z., Kalathil, D., and Ghavamzadeh, M. Robust reinforcement learning using offline data. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Panaganti, K., Xu, Z., Kalathil, D., and Ghavamzadeh, M. Bridging distributionally robust learning and offline rl: An approach to mitigate distribution shift and partial data coverage. *arXiv preprint arXiv:2310.18434*, 2023a.

Panaganti, K., Xu, Z., Kalathil, D., and Ghavamzadeh, M. Distributionally robust behavioral cloning for robust imitation learning. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 1342–1347, 2023b.

Pioch, N. J., Melhuish, J., Seidel, A., Santos Jr, E., Li, D., and Gorniak, M. Adversarial intent modeling using embedded simulation and temporal bayesian knowledge bases. In *Modeling and Simulation for Military Operations IV*, volume 7348, pp. 115–126, 2009.

Robey, A., Hassani, H., and Pappas, G. J. Model-based robust deep learning: Generalizing to natural, out-of-distribution data. *arXiv preprint arXiv:2005.10247*, 2020.

Rockafellar, R. T. and Wets, R. J.-B. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Russel, R. H. and Petrik, M. Beyond confidence regions: Tight bayesian ambiguity sets for robust mdps. *Advances in Neural Information Processing Systems*, 2019.

Scherrer, B., Ghavamzadeh, M., Gabillon, V., Lesner, B., and Geist, M. Approximate modified policy iteration and its application to the game of tetris. *J. Mach. Learn. Res.*, 16(49):1629–1676, 2015.

Schmidt, T., Hertkorn, K., Newcombe, R., Marton, Z., Suppa, M., and Fox, D. Depth-based tracking with physical constraints for robot manipulation. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 119–126, 2015.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shah, S., Dey, D., Lovett, C., and Kapoor, A. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics: Results of the 11th International Conference*, pp. 621–635. Springer, 2018.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shapiro, A. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017.

Shi, L. and Chi, Y. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *arXiv preprint arXiv:2208.05767*, 2022.

Shi, L., Li, G., Wei, Y., Chen, Y., Geist, M., and Chi, Y. The curious price of distributional robustness in reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 36, 2023.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Sinha, A., Namkoong, H., and Duchi, J. C. Certifiable distributional robustness with principled adversarial training. *ICLR, arXiv preprint arXiv:1710.10571*, 2018.

Song, Y., Zhou, Y., Sekhari, A., Bagnell, D., Krishnamurthy, A., and Sun, W. Hybrid rl: Using both offline and online data can make rl efficient. In *The Eleventh International Conference on Learning Representations*, 2023.

Sünderhauf, N., Brock, O., Scheirer, W., Hadsell, R., Fox, D., Leitner, J., Upcroft, B., Abbeel, P., Burgard, W., Milford, M., et al. The limits and potentials of deep learning for robotics. *The International journal of robotics research*, 37(4-5):405–420, 2018.

Szepesvári, C. and Munos, R. Finite time bounds for sampling based fitted value iteration. In *Proceedings of the 22nd international conference on Machine learning*, pp. 880–887, 2005.

Van Erven, T., Grunwald, P., Mehta, N. A., Reid, M., Williamson, R., et al. Fast rates in statistical and online learning. *JMLR*, 2015.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University press, 2018.

Wang, R., Foster, D., and Kakade, S. M. What are the statistical limits of offline {rl} with linear function approximation? In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=30EvkP2aQLD.

Wang, S., Si, N., Blanchet, J., and Zhou, Z. A finite sample complexity bound for distributionally robust q-learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3370–3398, 2023a.

Wang, S., Si, N., Blanchet, J., and Zhou, Z. Sample complexity of variance-reduced distributionally robust q-learning. *arXiv preprint arXiv:2305.18420*, 2023b.

Wang, Y. and Zou, S. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34:7193–7206, 2021.

Wang, Y. and Zou, S. Policy gradient method for robust reinforcement learning. In *International Conference on Machine Learning*, pp. 23484–23526, 2022.

Wang, Y., Hu, Y., Xiong, J., and Zou, S. Achieving minimax optimal sample complexity of offline reinforcement learning: A dro-based approach. *arXiv preprint arXiv:2305.13289v2*, 2023c.

Wiesemann, W., Kuhn, D., and Rustem, B. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34, 2021.

Xu, H. and Mannor, S. Distributionally robust Markov decision processes. In *Advances in Neural Information Processing Systems*, pp. 2505–2513, 2010.

Xu*, Z., Panaganti*, K., and Kalathil, D. Improved sample complexity bounds for distributionally robust reinforcement learning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Conference on Artificial Intelligence and Statistics, 2023.

Yang, J., Zhou, K., Li, Y., and Liu, Z. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

Yang, W., Wang, H., Kozuno, T., Jordan, S. M., and Zhang, Z. Avoiding model estimation in robust markov decision processes with a generative model. *arXiv preprint arXiv:2302.01248*, 2023.

Yu, P. and Xu, H. Distributionally robust counterpart in Markov decision processes. *IEEE Transactions on Automatic Control*, 61(9):2538–2543, 2015.

Zhang, R., Hu, Y., and Li, N. Regularized robust mdps and risk-sensitive mdps: Equivalence, policy gradient, and sample complexity. *ICLR, arXiv preprint arXiv:2306.11626*, 2024.

Zhou, R., Liu, T., Cheng, M., Kalathil, D., Kumar, P., and Tian, C. Natural actor-critic for robust reinforcement learning with function approximation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Supplementary Materials

A. Related Works

Offline RL: Offline RL tackles the problem of learning optimal policy using minimal amount of offline/historical data collected according to a behavior policy (Lange et al., 2012; Levine et al., 2020). Due to offline data quality and no access to simulators or any world models for exploration, the offline RL problem suffers from the out-of-distribution (Robey et al., 2020; Yang et al., 2021) challenge. Many works (Fujimoto et al., 2019; Kumar et al., 2019; 2020; Fujimoto & Gu, 2021; Kostrikov et al., 2021) have introduced deep offline RL algorithms aimed at alleviating the out-of-distribution issue by some variants of trust-region optimization (Schulman et al., 2015; 2017). The earliest and most promising theoretical investigations into model-free offline RL methodologies relied on the assumption of *uniformly bounded concentrability* such as the approximate modified policy iteration (AMPI) algorithm (Scherrer et al., 2015) and fitted Q-iteration (FQI) (Munos & Szepesvári, 2008) algorithm. This assumption mandates that the ratio of the state-action occupancy distribution induced by *any* policy to the data generating distribution remains uniformly bounded across all states and actions (Munos, 2007; Antos et al., 2008; Munos & Szepesvári, 2008; Farahmand et al., 2010; Chen & Jiang, 2019). This makes offline RL particularly challenging (Foster et al., 2022) and there have been efforts to understand the limits of this setting.

Robust RL: The robust Markov decision process framework (Nilim & El Ghaoui, 2005; Iyengar, 2005) tackles the challenge of formulating a policy resilient to model discrepancies between training and testing environments. Robust reinforcement learning problem pursues this objective in the data-driven domain. Deploying simplistic RL policies (Corporation, 2021) can lead to catastrophic outcomes when faced with evident disparities in models. The optimization techniques and analyses in robust RL draw inspiration from the distributionally robust optimization (DRO) toolkit in supervised learning (Duchi & Namkoong, 2018; Shapiro, 2017; Gao & Kleywegt, 2022; Bertsimas et al., 2018; Namkoong & Duchi, 2016; Blanchet et al., 2019). Many heuristic works (Xu & Mannor, 2010; Wiesemann et al., 2013; Yu & Xu, 2015; Mannor et al., 2016; Russel & Petrik, 2019) show robust RL is valuable in such scenarios involving disparities of a simulator model with the real-world model. Many recent works address fundamental issues of RMDP giving concrete theoretical understanding in terms of sample complexity (Panaganti & Kalathil, 2021b; 2022; Xu* et al., 2023; Shi & Chi, 2022; Shi et al., 2023). Many works (Panaganti & Kalathil, 2021a; Wang & Zou, 2021; Panaganti & Kalathil, 2022) devise model-free online and offline robust RL algorithms employing general function approximation to handle potentially infinite state spaces. Recent work (Panaganti et al., 2023b) introduces distributional robustness in the imitation learning setting. There have been works (Panaganti, 2023; Panaganti et al., 2023a; Wang et al., 2023c) connecting robust RL with offline RL by linking notions of robustness and pessimism.

B. Theoretical Discussions and Final Remarks

In this section, we first discuss the proof ideas for our results, focusing on discussions of the assumptions and their improvements. Next, we compare our results with the most relevant ones from the robust RL literature. Our Table 1 should be used as a reference. Finally, we discuss the bilinear model architecture in detail, as ours is the first work to consider it in the robust RL setting under the general function architecture for the value and dual functions approximations.

Discussions on Proof Sketch: We first discuss our RPQ algorithm (Algorithm 1) result. We note that the concentrability (Assumption 1) assumption requires the data-generating policy to be robust exploratory. That is, it covers the state-action occupancy induced by any policy and any φ-divergence set transition model. We reiterate the proof idea of the suboptimality result (Panaganti et al., 2022, Theorem 1) of the RFQI algorithm (Panaganti et al., 2022, Algorithm 1). We highlight the most important differences with Panaganti et al. (2022); Zhang et al. (2024) here. Firsty, we generalize the robust performance lemma ($\mathbb{E}_{s_0\sim d_0}[V^{\pi^*}] - \mathbb{E}_{s_0\sim d_0}[V^{\pi_K}] \leq 2\|Q^{\pi^*} - Q_K\|_{1,\nu}/(1-\gamma)$ at Equation (26)) for any general φ-divergence problem. Secondly, we identify that it is hard to come up with a unified analysis for general φ-divergences in robust RL setting via the dual reformulation of the distributionally robust optimization problem (Duchi & Namkoong, 2018, Proposition 1). Thus, a direct extension of the results in Panaganti et al. (2022) is hard for general φ-divergences. By RPQ analyses, we showcase that it is indeed possible to get a unified analysis for the robust RL problem using the RRMDP framework. Thirdly, we show the generalization bounds for the empirical risk minimization (Proposition 7) and least squares (Proposition 8) estimators for general φ-divergences with unified results. By these three points, equipped with the more general robust exploratory concentrability (Assumption 1), we have a unified general φ-divergences suboptimality result (Theorem 1) for the RPQ algorithm.

We now discuss our HyTQ algorithm (Algorithm 2) result. We immediately make an important note here. The concentrability assumption improvement is two-fold: all-policy concentrability (Assumption 9) to single concentrability, and then to the robust Bellman error transfer coefficient (Assumption 4) via Lemma 8. We refer to Foster et al. (2022); Song et al. (2023) for further discussion on such concentrability assumption improvements and tightness in the non-robust offline RL. We leave it to future work for more tightness of these assumptions in the robust RL setting. We execute a tighter analysis in our HyTQ algorithm result (Theorem 2) compared to our RPQ algorithm TV φ-divergence specialized result (Theorem 4). We summarize the steps as follows:
Step (a): We meticulously arrive at the following robust performance lemma (c.f. Equations (37) and (39)) for each algorithm iteration k: $\mathbb{E}_{s_0\sim d_0}[V_0^{\pi^*}(s_0)-V_0^{\pi_k}(s_0)] \leq \sum_{h=0}^{H-1}\mathbb{E}_{s,a\sim d_h^{\pi^*}}[(\mathcal{T}Q_{h+1}^k(s,a)-Q_h^k(s,a))_+]+\sum_{h=0}^{H-1}\mathbb{E}_{s,a\sim d_h^{\pi_k}}[(Q_h^k(s,a)-\mathcal{T}Q_{h+1}^k(s,a))_+]$. We highlight that the first summand here depends on the samples from state-action occupancy of the optimal robust policy and for the second summand it is the w.r.t. the learned HyTQ policies. It is now intuitive to connect the first summand with the offline samples and the second with the online samples.
Finally, step (b): With the above gathered intuition, firstly, the history dependent dataset collected by different offline data-generating policy and the learned HyTQ policies on the nominal model P^o warrants more sophisticated generalization bounds for the empirical risk minimization and least squares estimators. We prove a generalization bound for empirical risk minimization when the data are not necessarily i.i.d. but adapted to a stochastic process in Appendix D. This result is applicable to more machine learning problems outside of the scope of this paper as well. Finally, equipped with the transfer coefficient (Assumption 4) and bilinear model (Assumption 7) assumptions for the nominal model P^o, we formally show generalization bounds for the empirical risk minimization and least squares estimators in Propositions 9 and 10 respectively. We complete the proof by combining these two steps.

Remark 3. *We offer computational tractability in our RPQ and HyTQ algorithms due to the usage of empirical risk minimization (Steps 4 & 9 resp.), over the general function class $\mathcal{G}$, and least-squares (Steps 5 & 10 resp.), over the general function class $\mathcal{F}$,* computationally tractable *estimators. This two-step estimator update avoids the complexity of solving the inner problem for each state-action pair (leading to scaling issues for high-dimensional problems) in the original robust Bellman operators (Equations* (3) *and* (10)*). To the best of our knowledge, no purely online or purely offline robust RL algorithms are known to be tractable in this sense, except other robust Q-iteration and actor-critic methods (discussed in Section 1) and except under much stronger coverage conditions (like single-policy and uniform) in the tabular setting.*

Theoretical Guarantee Discussions: In the suboptimality result (Theorem 1) for the RPQ algorithm (Algorithm 1), we only mention the leading statistical bound with a problem-dependent (on φ-divergence) constant $c_\varphi(\lambda,\gamma)$. We provide the exact constants pertaining to different φ-divergences in a restated statement of Theorem 1 in Theorem 3. Furthermore, the constants c_1, c_2, c_3 in Theorem 3 take different values for different φ-divergences provided in Proposition 3. Similarly, for the suboptimality result (Theorem 2) of the HyTQ algorithm (Algorithm 2), we provide a more detailed bound in a restated statement in Theorem 5.

In the following we provide comparisons of suboptimality results with relevant prior works. But first, we make an important note here on ρ, the robustness radius parameter in RMDPs, and λ, the robustness penalization parameter in RRMDPs, mentioned briefly in Table 1. (Levy et al., 2020; Yang et al., 2023) establish the regularized and constrained versions of DRO and robust MDP problems, respectively, are equivalent by connecting their respective (λ and ρ) robustness parameters. Moreover, both observe rigorously that λ and ρ are inversely related. This is intuitively true, as $\lambda\to\infty$ and $\rho\to 0$ both yield the non-robust solutions on the nominal model P^o and as $\lambda\to 0$ and $\rho\to\infty$ both yield the conservative solutions considering the entire probability simplex for the transition dynamics. However, it is an interesting open problem to establish an exact analytical relation between the robustness parameters λ and ρ. We leave this to future research as it is out of the scope of this work.

Here we specialize our result (Theorem 3) for the chi-square φ-divergence $\gamma\mathrm{R}^3\mathrm{L}$ problem. We get the suboptimality for the RPQ algorithm as $\widetilde{\mathcal{O}}\left(\frac{\max\{\frac{1}{\lambda(1-\gamma)^2},\lambda\}\sqrt{C\log(|\mathcal{F}||\mathcal{G}|)}}{(1-\gamma)^2\sqrt{N}}\right)$, where we only have presented the higher-order terms. The suboptimality of Algorithm 2 in Yang et al. (2023, Theorem 5.1) for chi-square φ-divergence is stated for $\lambda=1/(1-\gamma)$ as $\widetilde{\mathcal{O}}\left(\frac{\max\{\frac{1}{(1-\gamma)^2},\sqrt{\log(|\mathcal{S}||\mathcal{A}|)}\}}{d_{\min}^3(1-\gamma)^3N^{1/3}}\right)$ where $d_{\min}$ is described in Table 1. We use the typical equivalence from RL literature for comparison between these two results in the tabular setting with generative/simulator modeling assumption: function approximation classes with full dimension yields $\log(|\mathcal{F}||\mathcal{G}|)=O(|\mathcal{S}||\mathcal{A}|)$ (Panaganti et al., 2022) and uniform support data sampling yields $\mu_{\min}=1/(|\mathcal{S}||\mathcal{A}|)$ and $C\leq|\mathcal{S}||\mathcal{A}|$ (Shi et al., 2023). Now our result with $\lambda=1/(1-\gamma)$ reduces

to $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\sqrt{N}}\right)$ and their result (Yang et al., 2023) reduces to $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|^3|\mathcal{A}|^3\max\{\frac{1}{(1-\gamma)^2},\sqrt{\log(|\mathcal{S}||\mathcal{A}|)}\}}{(1-\gamma)^3N^{1/3}}\right)$. Two comments warrant attention here. Firstly, compared to a model-based robust regularized algorithm (robust value iteration using empirical estimates of the nominal model P^o) (Yang et al., 2023, Theorem 3.2), our suboptimality bound is worse off by the factors $\sqrt{|\mathcal{S}||\mathcal{A}|}$ and $1/(1-\gamma)$. We leave it to future work to fine-tune and get optimal rates. Secondly, their result Yang et al. (2023, Theorem 5.1) exhibit inferior performance compared to ours in all parameters, but we do want to note that they make a first attempt to give suboptimality bounds for the stochastic approximation-based algorithm. The dependence on $|\mathcal{S}||\mathcal{A}|$ is typically known to be bad using the stochastic approximation technical tool (Chen et al., 2022), and Yang et al. (2023, Discussion on Page 16) conjectures using the Polyak-averaging technique to improve their suboptimality bound rate to $N^{-1/2}$.

Here we discuss and compare our result for the total variation φ-divergence setting. As mentioned in Remark 1, we have a specialized result in Appendix E.2 for the total variation φ-divergence. We get the suboptimality result (Theorem 4) for the RPQ algorithm as $\widetilde{\mathcal{O}}\left(\frac{\lambda\sqrt{C_{\text{tv}}\log(|\mathcal{F}||\mathcal{G}|)}}{(1-\gamma)^3\sqrt{N}}\right)$, where we again only have presented the higher-order terms. Panaganti et al. (2022, Theorem 1) mentioned in Table 1 also exhibits same suboptimality guarantee replacing λ with ρ^{-1}. As we noted before, ρ (the robustness radius parameter in RMDPs) and λ (the robustness penalization parameter in RRMDPs) are inversely related, and for the TV φ-divergence we observe a straightforward relation between the two as $\lambda=\rho^{-1}$. Using the earlier arguments for a tabular setting bound, our result further reduces to $\widetilde{\mathcal{O}}\left(\frac{\lambda|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\sqrt{N}}\right)$. Now comparing this to the minimax lower bound (Shi et al., 2023, Theorem 2), our suboptimality bound is worse off by the factors $\sqrt{|\mathcal{S}||\mathcal{A}|}$ and $1/(1-\gamma)$. Nevertheless, we push the boundaries by providing novel suboptimality guarantee studying the robust RL problem in the hybrid RL setting. Furthermore, as mentioned earlier in Remark 2, we provide the offline+online robust RL suboptimality guarantee $\widetilde{\mathcal{O}}\left(\max\{C(\pi^*),1\}\sqrt{dH^3}(\lambda+H)\log(|\mathcal{F}||\mathcal{G}|/\delta)/\sqrt{N}\right)$ in the Appendix F. We also remark that the HyTQ algorithm can be proposed under the RMDP setting with a similar suboptimality guarantee due to the similarity of the dual Bellman equations under the TV φ-divergence for RMDPs and RRMDPs (c.f. Equation (33) and Xu* et al. (2023, Lemma 8)). For the sake of consistency and novelty, we present our results solely for the RRMDP setting. As mentioned earlier, the concentrability assumption improvement is two-fold (Lemma 8): all-policy concentrability (Assumption 9) to single concentrability to transfer coefficient. This is the first of its kind result that does not yet have any existing lower bounds to compare in the robust RL setting. Under similar transfer coefficient, Bellman completeness, and bilinear model assumptions, the HyTQ algorithm sample complexity (Corollary 5) is comparable to that of a non-robust RL algorithm (Song et al., 2023), i.e., $\widetilde{\mathcal{O}}(\max\{(C(\pi^*))^2,1\}dH^5\log(H|\mathcal{F}|/\delta)/\varepsilon^2)$. We leave it to future work for developing minimax rates and getting optimal algorithm guarantees.

Here we specialize our result (Theorem 3) for the KL φ-divergence $\gamma\text{R}^3\text{L}$ problem. We get the suboptimality for RPQ as $\widetilde{\mathcal{O}}\left(\frac{(\lambda+(1-\gamma)^{-1})\exp\{(\lambda(1-\gamma))^{-1}\}\sqrt{C\log(|\mathcal{F}||\mathcal{G}|)}}{(1-\gamma)^2\sqrt{N}}\right)$, where we only have presented the higher-order terms. Using the earlier arguments for a tabular setting bound, our result with $\lambda=1/(1-\gamma)$ again reduces to $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\sqrt{N}}\right)$. Zhang et al. (2024, Theorem 5) mentioned in Table 1 also exhibits same suboptimality guarantee. Two remarks are in order here. Firstly, we remark that our RPQ algorithm and its theoretical guarantee unifies for a class of φ-divergence classes, whereas Zhang et al. (2024, Algorithm 1) is specialized for the KL φ-divergence. This steers towards our first main contribution discussed in Section 1. Secondly, we remark the robust regularized Bellman operator Equation (3) for the KL φ-divergence has a special form due to the existence of an analytical worse-case transition model. This arrives at a special structure of the form of an exponential robust Bellman operator in a Q-value-variant space. This special structure helps avoid the dual variable function update (Step 4) in the RPQ algorithm and the $\log(|\mathcal{G}|)$ factor in the suboptimal guarantee. We choose not to include this specialized result in this work (like we did for the TV φ-divergence in Appendix E.2) and directly point to Zhang et al. (2024). We do highlight here an important note for such a choice in our paper. The abovementioned special structure forces us to get online samples *from all the transition kernels* (c.f. Assumption 1), which is unrealistic in practice, to achieve an improvement in the hybrid robust RL setting. We leave it to future work for developing such improved algorithm guarantees in the hybrid robust RL setting for other φ-divergences.

Discussion of Bilinear Models in the Hybrid Robust RL setting: We emphasize that while our bilinear model for the HyTQ algorithm is specialized to low occupancy complexity (i.e. the occupancy measures themselves have a low-rank structure) and low-rank feature selection model (i.e. the nominal model P^o has a low-rank structure) in Appendix F.2, the function classes $\mathcal{F}$ (Q-value representations) and $\mathcal{G}$ (dual-value representations) can be arbitrary, potentially nonlinear

function classes (neural tangent kernels, neural networks, etc). Thus, even in the tabular setting with large state space (e.g. $|\mathcal{S}| > O(10^5)$) for the bilinear model, our suboptimality bounds only scale with the complexity of the function classes $\mathcal{F}$ and $\mathcal{G}$, which can considerably be low compared to $|\mathcal{S}|$. For example, linear function approximators (e.g. linear feature dimension $d = \log(|\mathcal{F}||\mathcal{G}|) \ll |\mathcal{S}||\mathcal{A}|$), RKHS approximators with low dimension features, neural tangent kernels with low effective neural net dimension, and more function approximators. Moreover, our work solves the robust RL problem with more nuances, which is at least as hard as the non-robust RL problem. Thus, due to the new upcoming research status of robust RL in the general function approximation setting, we believe it is currently out of scope for this work to satisfy more general bilinear model classes (Du et al., 2021). Nevertheless, our initial findings for robust RL by the HyTQ algorithm in the hybrid learning setting reveal the hardness of finding larger model classes for RRMDPs with general φ-divergences.

C. Useful Technical Results

We state the following result from the *penalized distributionally robust optimization* literature (Levy et al., 2020).

Lemma 1 (Levy et al., 2020, Section A.1.2)**.** *Let P^o be a distribution on the space $\mathcal{X}$ and let $l : \mathcal{X} \to \mathbb{R}$ be a loss function. For φ-divergence (1), we have*

$$\sup_{P \ll P^o} \mathbb{E}_P[l(X) - \lambda D_\varphi(P, P^o)] = \inf_{\eta \in \mathbb{R}} \ \lambda \mathbb{E}_{P^o}\left[\varphi^*\left(\frac{l(X) - \eta}{\lambda}\right)\right] + \eta,$$

where $\varphi^*(s) = \sup_{t \geq 0}\{st - \varphi(t)\}$ is the Fenchel conjugate function of φ. Moreover, the optimization on the right hand side is convex in η.

We state a standard concentration inequality here.

Lemma 2 (Bernstein's Inequality (Vershynin, 2018, Theorem 2.8.4))**.** *Fix any $\delta \in (0, 1)$. If $X_1, \cdots, X_T$ are independent and identically distributed random variables with finite second moment. Assume that $|X_t - \mathbb{E}[X_t]| \leq M$, for all t. Then we have with probability at least $1 - \delta$:*

$$\left|\mathbb{E}[X_1] - \frac{1}{T}\sum_{t=1}^{T} X_t\right| \leq \sqrt{\frac{2\mathbb{E}[X_1^2]\log(2/\delta)}{T}} + \frac{M\log(2/\delta)}{3T}.$$

We now state a useful concentration inequality when the samples are not necessarily i.i.d. but adapted to a stochastic process.

Lemma 3 (Freedman's Inequality (Song et al., 2023, Lemma 14))**.** *Let $X_1, \cdots, X_T$ be a sequence of $M > 0$-bounded real valued random variables where $X_t \sim P_t$ from some stochastic process P_t that depends on the history $X_1, \cdots, X_{t-1}$. Then, for any $\delta > 0$ and $\lambda \in [0, 1/2M]$, we have with probability at least $1 - \delta$:*

$$\left|\sum_{t=1}^{T}(X_t - \mathbb{E}[X_t \mid P_t])\right| \leq \lambda \sum_{t=1}^{T}(2M|\mathbb{E}[X_t \mid P_t]| + \mathbb{E}[X_t^2 \mid P_t]) + \frac{\log(2/\delta)}{\lambda}.$$

We now state a result for the generalization bounds on empirical risk minimization (ERM) (Shalev-Shwartz & Ben-David, 2014).

Lemma 4 (ERM Generalization Bound (Panaganti et al., 2022, Lemma 3))**.** *Let P be the data generating distribution on the space $\mathcal{X}$ and let $\mathcal{H}$ be a given hypothesis class of functions. Assume that for all $x \in \mathcal{X}$ and $h \in \mathcal{H}$ for loss function l we have that $|l(h, x)| \leq c_1$ for some positive constant $c_1 > 0$ and $l(h, x)$ is c_3-Lipschitz in h. Given a dataset $\mathcal{D} = \{X_i\}_{i=1}^N$, generated independently from P, denote $\hat{h}$ as the ERM solution, i.e. $\hat{h} = \arg\min_{h \in \mathcal{H}}(1/N)\sum_{i=1}^N l(h, X_i)$. Furthermore, let $\mathcal{H}$ be a finite hypothesis class, i.e. $|\mathcal{H}| < \infty$, with $|h \circ x| \leq c_2$ for all $h \in \mathcal{H}$ and $x \in \mathcal{X}$. For any fixed $\delta \in (0, 1)$ and $h^* \in \arg\min_{h \in \mathcal{H}} \mathbb{E}_{X \sim P}[l(h, X)]$, we have*

$$\mathbb{E}_{X \sim P}[l(\hat{h}, X)] - \mathbb{E}_{X \sim P}[l(h^*, X)] \leq 2c_2c_3\sqrt{\frac{2\log(|\mathcal{H}|)}{N}} + 5c_1\sqrt{\frac{2\log(8/\delta)}{N}},$$

with probability at least $1 - \delta$.

We now state a result from variational analysis literature (Rockafellar & Wets, 2009) that is useful to relate minimization of integrals and the integrals of pointwise minimization under decomposable spaces.

Remark 4. *A few examples of decomposable spaces are $L^p(\mathcal{S}\times\mathcal{A}, \Sigma(\mathcal{S}\times\mathcal{A}), \mu)$, for any $p \geq 1$, and $\mathcal{M}(\mathcal{S}\times\mathcal{A}, \Sigma(\mathcal{S}\times\mathcal{A}))$, the space of all $\Sigma(\mathcal{S}\times\mathcal{A})$-measurable functions.*

Lemma 5 (Rockafellar & Wets, 2009, Theorem 14.60)**.** *Let $\mathcal{X}$ be a space of measurable functions from Ω to $\mathbb{R}$ that is decomposable relative to a σ-finite measure μ on the σ-algebra $\mathcal{A}$. Let $f : \Omega\times\mathbb{R} \to \mathbb{R}$ (finite-valued) be a normal integrand. Then, we have*

$$\inf_{x\in\mathcal{X}} \int_{\omega\in\Omega} f(\omega, x(\omega))\mu(\mathrm{d}\omega) = \int_{\omega\in\Omega} \left(\inf_{x\in\mathbb{R}} f(\omega, x)\right) \mu(\mathrm{d}\omega).$$

Moreover, as long as the above infimum is finite, we have that $x' \in \arg\min_{x\in\mathcal{X}} \int_{\omega\in\Omega} f(\omega, x(\omega))\mu(\mathrm{d}\omega)$ if and only if $x'(\omega) \in \arg\min_{x\in\mathbb{R}} f(\omega, x)$ for μ-almost everywhere.

Now we state a few results that will be useful for the analysis of our finite-horizon results in this work. The following result (Song et al., 2023, Lemma 6) is useful under the use of bilinear model approximation. This result follows from the elliptical potential lemma (Lattimore & Szepesvári, 2020, Lemma 19.4) for deterministic vectors.

Lemma 6 (Elliptical Potential Lemma)**.** *Let $X_h(f^1), \cdots, X_h(f^T) \in \mathbb{R}^d$ be a sequence of vectors with $\|X_h(f^t)\| \leq B_X < \infty$ for all $t \leq T$ and fix $\sigma \geq B_X^2$. Define $\Sigma_{t;h} = \sum_{\tau=1}^{t} X_h(f^\tau)X_h(f^\tau)^\top + \sigma\mathbb{1}_{d\times d}$ for $t \in [T]$. Then, the following holds: $\sum_{t=1}^{T} \|X_h(f^t)\|_{\Sigma_{t-1;h}^{-1}} \leq \sqrt{2dT\log(1 + (TB_X^2/(\sigma d)))}$.*

We now state a result for the generalization bounds on the least-squares regression problem when the data are not necessarily i.i.d. but adapted to a stochastic process. We refer to Van Erven et al. (2015) for more statistical and online learning generalization bounds for a wider class of loss functions.

Lemma 7 (Online Least-squares Generalization Bound (Song et al., 2023, Lemma 3))**.** *Let $L, M > 0$, $\delta \in (0, 1)$, and let $\mathcal{X}$ be an input space and $\mathcal{Y}$ be a target space . Let $\mathcal{H} : \mathcal{X} \mapsto [-M, M]$ be a given real-valued hypothesis class of functions with $|\mathcal{H}| < \infty$. Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, denote $\widehat{h}$ as the least square solution, i.e. $\widehat{h} = \arg\min_{h\in\mathcal{H}} \sum_{i=1}^N (h(x_t) - y_t)^2$. The dataset $\mathcal{D}$ is generated as $x_t \sim P_t$ from some stochastic process P_t that depends on the history $\{(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})\}$, and y_t is sampled via the conditional probability $p(\cdot \mid x_t)$ as $y_t \sim p(\cdot \mid x_t) = h^*(x_t) + \varepsilon_t$, where the function h^* satisfies approximate realizability i.e. $\inf_{h\in\mathcal{H}}(1/N)\sum_{t=1}^N \mathbb{E}_{x\sim P_t}(h^*(x) - h(x))^2 \leq \gamma$, and $\varepsilon_t{}_{t=1}^N$ are independent random variables such that $\mathbb{E}[y_t \mid x_t] = h^*(x_t)$. Suppose it also holds $\max_t |y_t| \leq L$ and $\max_x |h^*(x)| \leq M$. Then, the least square solution satisfies with probability at least $1 - \delta$:*

$$\sum_{t=1}^{N} \mathbb{E}_{x\sim P_t}(\widehat{h}(x) - h^*(x))^2 \leq 3\gamma N + 64(L + M)^2 \log(2|\mathcal{H}|/\delta).$$

D. Useful Foundational Results

We provide the following result highlighting the necessary characteristics for specific examples of the Fenchel conjugate functions φ^*.

Proposition 3 (φ-Divergence Bounds)**.** *Let $V \in [0, V_{\max}]^{|\mathcal{S}|}$ be any value function and fix a probability distribution $P^o \in \Delta(\mathcal{S})$. Define $h(y, \eta) = (\lambda\varphi^*((\eta - y)/\lambda) - \eta)$. Consider the following scalar convex optimization problem: $\inf_{\eta\in\Theta\subseteq\mathbb{R}} \mathbb{E}_{s\sim P^o} h(V(s), \eta)$. Let the maximum absolute value in Θ be less than or equal to c_3, let $|h(V(s), \eta)| \leq c_1$ for all $\eta \in \Theta$, and let $h(V(s), \eta)$ be c_2-Lipschitz in η; hold for some positive constants c_1, c_2, c_3. We have the following results for different forms of φ:*
(i) Let Assumption 8 hold. For TV distance i.e. $\varphi(t) = |t-1|/2$, we have $\Theta \equiv [-\lambda/2, \lambda/2]$, hence $c_3 = \lambda/2$, $c_1 = 2\lambda + V_{\max}$, and $c_2 = 2$.
(ii) For chi-square divergence i.e. $\varphi(t) = (t - 1)^2$, we have $\Theta \equiv [-\lambda, 2V_{\max} + 2\lambda]$, hence $c_3 = 2V_{\max} + 2\lambda$, $c_1 = \lambda + (2V_{\max} + 4\lambda)(\frac{2V_{\max}}{4\lambda} + 2)$, and $c_2 = (3 + \frac{V_{\max}}{\lambda})$.
(iii) For KL divergence i.e. $\varphi(t) = (t - 1)^2$, we have $\Theta \equiv [\lambda, V_{\max} + \lambda]$, hence $c_3 = V_{\max} + \lambda$, $c_1 = \lambda(\exp(\frac{V_{\max}}{\lambda}) - 1)$, and $c_2 = (\exp(\frac{V_{\max}}{\lambda}) + 1)$.
(iv) Fix $\alpha \in (0, 1)$. For α-CVaR i.e. $\varphi(t) = \mathbb{1}[0, 1/\alpha)$, we have $\Theta \equiv [0, V_{\max}/(1 - \alpha)]$, hence $c_3 = V_{\max}/(1 - \alpha)$, $c_1 = 2V_{\max}/(\alpha(1 - \alpha))$, and $c_2 = 1 + \alpha^{-1}$.

Proof. We first prove the statement for TV distance with $\varphi(t) = |t - 1|/2$. From φ-divergence literature (Xu* et al., 2023),

we know

$$\varphi^*(s) = \begin{cases} -\frac{1}{2} & s \le -\frac{1}{2}, \\ s & s \in [-\frac{1}{2}, \frac{1}{2}] \\ +\infty & s > \frac{1}{2}. \end{cases}.$$

Thus, we have

$$\begin{aligned}
\inf_{\eta\in\mathbb{R}} \mathbb{E}_{s\sim P^o} h(V(s),\eta) &= \inf_{\eta\in\mathbb{R}} \mathbb{E}_{s\sim P^o}[\lambda\varphi^*(\frac{\eta - V(s)}{\lambda})] - \eta \\
&\overset{(a)}{=} \inf_{\eta\in\mathbb{R},\, \frac{\eta-\min_s V(s)}{\lambda}\le\frac{1}{2}} \mathbb{E}_{s\sim P^o}[\lambda \max\{\frac{\eta - V(s)}{\lambda}, -\frac{1}{2}\}] - \eta \\
&\overset{(b)}{=} \inf_{\eta\in\mathbb{R},\, \frac{\eta}{\lambda}\le\frac{1}{2}} \mathbb{E}_{s\sim P^o}[\lambda \max\{\frac{\eta - V(s)}{\lambda}, -\frac{1}{2}\}] - \eta \\
&\overset{(c)}{=} \inf_{\eta\in\mathbb{R},\, \frac{\eta}{\lambda}\le\frac{1}{2}} \mathbb{E}_{s\sim P^o}[(\eta - V(s) + \lambda/2)_+] - \lambda/2 - \eta \\
&\overset{(d)}{=} \inf_{\eta'\in\mathbb{R},\eta'\le\lambda} \mathbb{E}_{s\sim P^o}[(\eta' - V(s))_+] - \eta' \\
&\overset{(e)}{=} \inf_{0\le\eta'\le\lambda} \mathbb{E}_{s\sim P^o}[(\eta' - V(s))_+] - \eta',
\end{aligned} \tag{17}$$

where (a) follows by definition of φ^*, (b) by Assumption 8, (c) by the fact $\max\{x,y\} = (x-y)_+ + y$ for any $x, y \in \mathbb{R}$, and (d) follows by making the substitution $\eta = \eta' - \lambda/2$. Finally, for (e), notice that since $V(s) \ge 0$, $\mathbb{E}_{s\sim P^o}[(\eta' - V(s))_+] - \eta' = -\eta' \ge 0$ holds when $\eta' \le 0$. So $\inf_{\eta'\in(-\infty,0]} \mathbb{E}_{s\sim P^o}[(\eta' - V(s))_+] - \eta' = 0$ is achieved at $\eta' = 0$.

We immediately have $\Theta \equiv [-\lambda/2, \lambda/2]$ since $\eta = \eta' - \lambda/2$. Since $\eta \le \lambda/2$ and $V(s) \le V_{\max}$, we further get $|h(V(s),\eta)| \le 2\lambda + V_{\max}$. For $\eta_1, \eta_2 \in \Theta$, from the fact $|(x)_+ - (y)_+| \le |(x-y)_+| \le |x-y|$ we have $|h(V(s),\eta_1) - h(V(s),\eta_2)| \le 2|\eta_1 - \eta_2|$. This proves statement (i).

We now prove the statement for chi-square divergence with $\varphi(t) = (t-1)^2$ following similar steps as before. From φ-divergence literature (Xu* et al., 2023), we know $\varphi^*(s) = (s/2+1)_+^2 - 1$. Thus, we have

$$\begin{aligned}
\inf_{\eta\in\mathbb{R}} \mathbb{E}_{s\sim P^o} h(V(s),\eta) &= \inf_{\eta\in\mathbb{R}} \mathbb{E}_{s\sim P^o}[\lambda\varphi^*(\frac{\eta - V(s)}{\lambda})] - \eta \\
&= \inf_{\eta\in\mathbb{R}} \mathbb{E}_{s\sim P^o}[\lambda(\frac{\eta - V(s)}{2\lambda} + 1)_+^2] - \lambda - \eta \\
&\overset{(f)}{=} \inf_{\eta'\in\mathbb{R}} \frac{1}{4\lambda}\mathbb{E}_{s\sim P^o}[(\eta' - V(s))_+^2] + \lambda - \eta' \\
&\overset{(g)}{=} \inf_{\eta'\in[\lambda, 2V_{\max}+4\lambda]} \frac{1}{4\lambda}\mathbb{E}_{s\sim P^o}[(\eta' - V(s))_+^2] + \lambda - \eta',
\end{aligned}$$

where (f) follows by making the substitution $\eta = \eta' - 2\lambda$. Finally, for (g), observe that the function $g(\eta') = \frac{1}{4\lambda}\mathbb{E}_{s\sim P^o}[(\eta' - V(s))_+^2] + \lambda - \eta'$ is convex in the dual variable η' and $\inf_{\eta'\in\mathbb{R}} g(\eta') \le 0$ since it is a Lagrangian dual variable. Since $V(s) \ge 0$, $\lambda - \eta'_* \le 0$ where η'_* is any solution of $\inf_{\eta'\in\mathbb{R}} g(\eta') \le 0$. When $\eta' \ge 2V_{\max} + 4\lambda$, notice that $g(\eta') \ge \frac{1}{4\lambda}(\eta'^2 - 2(V_{\max} + 2\lambda)\eta' + 4\lambda^2) \ge \lambda > 0$, since $0 \le V(s) \le V_{\max}$.

We immediately have $\Theta \equiv [-\lambda, 2V_{\max} + 2\lambda]$ since $\eta = \eta' - 2\lambda$. Since $\eta \le 2V_{\max} + 2\lambda$ and $V(s) \ge 0$, we further get $|h(V(s),\eta)| \le \lambda + (2V_{\max} + 4\lambda)(\frac{2V_{\max}}{4\lambda} + 2)$. For $\eta_1, \eta_2 \in \Theta$, from the facts $|(x)_+ - (y)_+| \le |(x-y)_+| \le |x-y|$ and $|(x)_+^2 - (y)_+^2| = |(x)_+ - (y)_+|((x)_+ + (y)_+)$, we have $|h(V(s),\eta_1) - h(V(s),\eta_2)| \le (3 + (V_{\max}))|\eta_1 - \eta_2|$. This proves statement (ii).

We now prove the statement for KL divergence with $\varphi(t) = t\log t$ following similar steps as before. From φ-divergence literature (Xu* et al., 2023), we know $\varphi^*(s) = \exp(s-1)$. Thus, we have

$$\inf_{\eta\in\mathbb{R}} \mathbb{E}_{s\sim P^o} h(V(s),\eta) = \inf_{\eta\in\mathbb{R}} \mathbb{E}_{s\sim P^o}[\lambda\varphi^*(\frac{\eta - V(s)}{\lambda})] - \eta$$

$$
\begin{aligned}
&= \inf_{\eta\in\mathbb{R}} \mathbb{E}_{s\sim P^o}[\lambda \exp(\frac{\eta - V(s)}{\lambda} - 1)] - \eta \\
&\overset{(h)}{=} \inf_{\eta'\in\mathbb{R}} \lambda\mathbb{E}_{s\sim P^o}[\exp(\frac{-\eta' - V(s)}{\lambda} - 1)] + \eta' \\
&\overset{(j)}{=} \inf_{\eta'\in[-\lambda-V_{\max},-\lambda]} \lambda\mathbb{E}_{s\sim P^o}[\exp(\frac{-\eta' - V(s)}{\lambda} - 1)] + \eta',
\end{aligned}
$$

where (h) follows by making the substitution $\eta = -\eta'$. Finally, for (j), observe that the function $g(\eta') = \lambda\mathbb{E}_{s\sim P^o}[\exp(\frac{-\eta'-V(s)}{\lambda} - 1)] + \eta'$ is convex in the dual variable η' since it is a Lagrangian dual variable. From Calculus, the optimal $\eta' = -\lambda + \lambda \log \mathbb{E}_{s\sim P^o} \exp(-V(s)/\lambda)$. So $\eta' \in [-\lambda - V_{\max}, -\lambda]$ since $0 \le V(s) \le V_{\max}$.

We immediately have $\Theta \equiv [\lambda, V_{\max} + \lambda]$ since $\eta = -\eta'$. Since $\eta \le V_{\max} + \lambda$ and $V(s) \ge 0$, we further get $|h(V(s), \eta)| \le \lambda(\exp(\frac{V_{\max}}{\lambda}) - 1)$. For $\eta_1, \eta_2 \in \Theta$, from the fact $\exp(-x)$ is 1-Lipschitz for $x \ge 0$, we have $|h(V(s), \eta_1) - h(V(s), \eta_2)| \le (\exp(\frac{V_{\max}}{\lambda}) + 1)|\eta_1 - \eta_2|$. This proves statement (ii).

We now prove the statement for α-CVAR with $\varphi(t) = \mathbb{1}[0, 1/\alpha)$. From φ-divergence literature (Levy et al., 2020), we know $\varphi^*(s) = (s)_+/\alpha$. Thus, we have

$$
\begin{aligned}
\inf_{\eta\in\mathbb{R}} \mathbb{E}_{s\sim P^o} h(V(s), \eta) &= \inf_{\eta\in\mathbb{R}} \mathbb{E}_{s\sim P^o}[\lambda\varphi^*(\frac{\eta - V(s)}{\lambda})] - \eta \\
&= \inf_{\eta\in\mathbb{R}} \frac{1}{\alpha}\mathbb{E}_{s\sim P^o}[(\eta - V(s))_+] - \eta \\
&\overset{(k)}{=} \inf_{0\le\eta\le V_{\max}/(1-\alpha)} \frac{1}{\alpha}\mathbb{E}_{s\sim P^o}[(\eta - V(s))_+] - \eta. \qquad (18)
\end{aligned}
$$

For (k), notice that since $V(s) \ge 0$, $(1/\alpha)\mathbb{E}_{s\sim P^o}[(\eta - V(s))_+] - \eta = -\eta \ge 0$ holds when $\eta \le 0$. Also, since $V(s) \le V_{\max}$, $(1/\alpha)\mathbb{E}_{s\sim P^o}[(\eta - V(s))_+] - \eta \ge 0$ holds when $\eta \ge V_{\max}/(1-\alpha)$.

We immediately have $\Theta \equiv [0, V_{\max}/(1-\alpha)]$. We further get $|h(V(s), \eta)| \le 2V_{\max}/(\alpha(1-\alpha))$. For $\eta_1, \eta_2 \in \Theta$, from the fact $|(x)_+ - (y)_+| \le |(x-y)_+| \le |x-y|$ we have $|h(V(s), \eta_1) - h(V(s), \eta_2)| \le (1 + \alpha^{-1})|\eta_1 - \eta_2|$. This proves the final statement of this result. □

We now state and prove a generalization bound for empirical risk minimization when the data are not necessarily i.i.d. but adapted to a stochastic process. This result is of independent interest to more machine learning problems outside of the scope of this paper as well. Furthermore, this result showcases better rate dependence on N, from $\widetilde{\mathcal{O}}(1/\sqrt{N})$ to $\widetilde{\mathcal{O}}(1/N)$, than the classical result Lemma 4 (Shalev-Shwartz & Ben-David, 2014). This result is not surprising and we refer to Van Erven et al. (2015, Theorems 7.6 & 5.4), in the i.i.d. setting, for such $\widetilde{\mathcal{O}}(1/N)$ fast rates with bounded losses to empirical risk minimization and beyond.

Proposition 4 (Online ERM Generalization Bound)**.** *Let $N > 0$, $\delta \in (0, 1)$, let $\mathcal{X}$ be an input space, and let $\mathcal{Y}$ be the target functional space. Let $\mathcal{H} \subseteq \mathcal{Y}$ be the given finite class of functions. Assume that for all $x \in \mathcal{X}$ and $h \in \mathcal{H}$ for loss function l we have that $|l(h(x))| \le c$ for some positive constant $c > 0$. Given a dataset $\mathcal{D} = \{x_i\}_{i=1}^N$, denote $\widehat{h}$ as the ERM solution, i.e. $\widehat{h} \leftarrow \arg\min_{h\in\mathcal{H}} \sum_{i=1}^N l(h(x_i))$. The dataset $\mathcal{D}$ is generated as $x_t \sim P_t$ from some stochastic process P_t that depends on the history $\{x_1, \ldots, x_{t-1}\}$, where the function $h_t^* \in \arg\min_{f\in\mathcal{Y}} \mathbb{E}_{x\sim P_t}[l(f(x))]$ satisfies approximate realizability i.e.*

$$
\inf_{h\in\mathcal{H}} \frac{1}{N} \sum_{t=1}^N \mathbb{E}_{x_t\sim P_t}(l(h(x_t)) - l(h_t^*(x_t))) \le \gamma,
$$

and for all $x \in \mathcal{X}$, $|l(h_t^*(x))| \le c$. Then, the ERM solution satisfies

$$
\sum_{t=1}^N \mathbb{E}_{x_t\sim P_t} l(\widehat{h}(x_t)) - \sum_{t=1}^N \mathbb{E}_{x_t\sim P_t} l(h_t^*(x_t)) \le 3\gamma N + 48c \log(2\,|\mathcal{H}|\,/\delta)
$$

with probability at least $1 - \delta$.

Proof. We adapt the proof of least-squares generalization bound (Song et al., 2023, Lemma 3) here for the empirical risk minimization generalization bound under online data collection. Fix any function $h \in \mathcal{H}$. We define the random variable $Z_t^h = l(h(x_t)) - l(h_t^*(x_t))$. Immediately, we note $|Z_t^h| \leq 2c$ for all t. By definition of h_t^*, we have a non-negative first moment of Z_t^h:

$$\mathbb{E}_{P_t}[Z_t^h] = \mathbb{E}_{x_t \sim P_t} l(h(x_t)) - \mathbb{E}_{x_t \sim P_t} l(h_t^*(x_t)). \tag{19}$$

By symmetrization, assuming $l(h_t^*(x_t))^2 \leq l(h(x_t))^2$, we have that

$$\begin{aligned} 0 \leq \mathbb{E}_{P_t}[(Z_t^h)^2] &\leq \mathbb{E}_{x_t \sim P_t}[2l(h(x_t))^2 - 2 \cdot l(h(x_t)) \cdot l(h_t^*(x_t))] \\ &\leq 2|l(h(x_t))| \mathbb{E}_{x_t \sim P_t}(l(h(x_t)) - l(h_t^*(x_t))) \\ &\leq 2c \cdot \mathbb{E}_{x_t \sim P_t}(l(h(x_t)) - l(h_t^*(x_t))). \end{aligned}$$

Similarly assuming $l(h_t^*(x_t))^2 \geq l(h(x_t))^2$, we get $0 \leq \mathbb{E}_{P_t}[(Z_t^h)^2] \leq 2c \cdot \mathbb{E}_{x_t \sim P_t}(l(h(x_t)) - l(h_t^*(x_t)))$. Thus, uniformly, we have

$$0 \leq \mathbb{E}_{P_t}[(Z_t^h)^2] \leq 2c \cdot \mathbb{E}_{x_t \sim P_t}(l(h(x_t)) - l(h_t^*(x_t))). \tag{20}$$

We remark that (20) is called *Bernstein condition* (Van Erven et al., 2015, Definition 5.1) when all sampling distributions P_t's are identical. This is one of the sufficient conditions on the loss functions to get $\mathcal{O}(1/N)$-generalization bounds for empirical risk minimization.

Now, applying Lemma 3 with $\lambda \in [0, 1/4c]$ and $\delta > 0$, we have

$$\begin{aligned} \left| \sum_{t=1}^N Z_t^h - \mathbb{E}_{P_t}[Z_t^h] \right| &\leq \lambda \sum_{t=1}^N (4c|\mathbb{E}_{P_t}[Z_t^h]| + \mathbb{E}_{P_t}[(Z_t^h)^2]) + \frac{\log(2/\delta)}{\lambda} \\ &\leq 6c\lambda \sum_{t=1}^N \mathbb{E}_{x_t \sim P_t}(l(h(x_t)) - l(h_t^*(x_t))) + \frac{\log(2/\delta)}{\lambda} \end{aligned}$$

with probability at least $1 - \delta$, where the last inequality uses (19) and (20). We set $\lambda = 1/12c$ in the above, we get for any $h \in \mathcal{H}$, with probability at least $1 - \delta$:

$$\left| \sum_{t=1}^N Z_t^h - \mathbb{E}_{P_t}[Z_t^h] \right| \leq \frac{1}{2} \sum_{t=1}^N \mathbb{E}_{x_t \sim P_t}(l(h(x_t)) - l(h_t^*(x_t))) + 12c \log(2\,|\mathcal{H}|\,/\delta),$$

by union bound over $h \in \mathcal{H}$. Using (19), we rearrange the above to get:

$$\sum_{t=1}^N Z_t^h \leq \frac{3}{2} \sum_{t=1}^N \mathbb{E}_{x_t \sim P_t}(l(h(x_t)) - l(h_t^*(x_t))) + 12c \log(2\,|\mathcal{H}|\,/\delta) \tag{21}$$

and

$$\sum_{t=1}^N \mathbb{E}_{x_t \sim P_t}(l(h(x_t)) - l(h_t^*(x_t))) \leq 2 \sum_{t=1}^N Z_t^h + 24c \log(2\,|\mathcal{H}|\,/\delta). \tag{22}$$

Define the function $\widetilde{h} \in \arg\min_{h \in \mathcal{H}} \sum_{t=1}^N \mathbb{E}_{x_t \sim P_t}(l(h(x_t)) - l(h_t^*(x_t)))$, which is independent of the dataset $\mathcal{D}$. By (21) for $\widetilde{h}$ and the approximate realizability assumption, we get

$$\sum_{t=1}^N Z_t^{\widetilde{h}} \leq \frac{3}{2} \sum_{t=1}^N \mathbb{E}_{x_t \sim P_t}(l(h(x_t)) - l(h_t^*(x_t))) + 12c \log(2\,|\mathcal{H}|\,/\delta) \leq \frac{3}{2}\gamma N + 12c \log(2\,|\mathcal{H}|\,/\delta).$$

By definitions of $\widetilde{h}$ and the ERM function $\widehat{h}$, we have that

$$\sum_{t=1}^N Z_t^{\widehat{h}} = \sum_{t=1}^N l(\widehat{h}(x_t)) - l(h_t^*(x_t)) \leq \sum_{t=1}^N l(\widetilde{h}(x_t)) - l(h_t^*(x_t)) = \sum_{t=1}^N Z_t^{\widetilde{h}}.$$

From the above two relations, we get

$$\sum_{t=1}^{N} Z_t^{\widehat{h}} \leq \frac{3}{2}\gamma N + 12c\log(2\left|\mathcal{H}\right|/\delta).$$

Now, using this and using (22) for the function $\widehat{h}$, we get

$$\sum_{t=1}^{N} \mathbb{E}_{x_t\sim P_t} l(\widehat{h}(x_t)) - l(h_t^*(x_t)) \leq 2\sum_{t=1}^{N} Z_t^{\widehat{h}} + 24c\log(2\left|\mathcal{H}\right|/\delta) \leq 3\gamma N + 48c\log(2\left|\mathcal{H}\right|/\delta),$$

which holds with probability at least $1-\delta$. This completes the proof. □

We now state a useful result for an infinite-horizon discounted robust φ-regularized Markov decision process $(\mathcal{S}, \mathcal{A}, r, P^o, \lambda, \gamma, \varphi, d_0)$. This result helps our RPQ algorithm's policy search space to be the class of deterministic Markov policies.

Proposition 5. *The robust regularized Bellman operator $\mathcal{T}$ (3)*

$$(\mathcal{T}Q)(s,a) = r(s,a) + \gamma \inf_{P_{s,a}\in\mathcal{P}_{s,a}} \left(\mathbb{E}_{s'\sim P_{s,a}}[\max_{a'} Q(s',a')] + \lambda D_\varphi(P_{s,a}, P^o_{s,a})\right),$$

and the value function operator $(\mathcal{T}_v V)(\cdot) = \max_a(\mathcal{T}Q)(\cdot,a)$ are both γ-contraction operators w.r.t sup-norm. Moreover, their respective unique fixed points Q^*_λ and V^*_λ, for optimal policy π^*, achieve the optimal robust value $\max_\pi V^\pi_\lambda$. Furthermore, the robust regularized optimal policy π^* is a deterministic Markov policy satisfying $\pi^*(\cdot) = \arg\max_a Q^*_\lambda(\cdot,a)$.

Proof. The γ-contraction property of both operators directly follow from the fact $\inf_x p(x) - \inf_x q(x) \leq \sup_x(p(x)-q(x))$. Furthermore, this result is a direct corollary of (Yang et al., 2023, Proposition 3.1) and (Iyengar, 2005, Corollary 3.1). □

We now state a similar result for a finite-horizon discounted robust φ-regularized Markov decision process $(\mathcal{S}, \mathcal{A}, P^o = (P^o_h)_{h=0}^{H-1}, r = (r_h)_{h=0}^{H-1}, \lambda, H, \varphi, d_0)$. This result helps our HyTQ algorithm's policy search space to be the class of non-stationary deterministic Markov policies.

Proposition 6. *The robust regularized Bellman operator $\mathcal{T}$ (10) and the value function operator $\mathcal{T}_v$ are as follows:*

$$(\mathcal{T}Q_{h+1})(s,a) = r_h(s,a) + \inf_{P_{h,s,a}\in\mathcal{P}_{h,s,a}} \left(\mathbb{E}_{s'\sim P_{h,s,a}}[\max_{a'} Q_{h+1}(s',a')] + \lambda D_\varphi(P_{h,s,a}, P^o_{h,s,a})\right) \quad \textit{and}$$

$$(\mathcal{T}_v V_{h+1})(s) = \max_a \left[r_h(s,a) + \inf_{P_{h,s,a}\in\mathcal{P}_{h,s,a}} \left(\mathbb{E}_{s'\sim P_{h,s,a}}[V_{h+1}(s')] + \lambda D_\varphi(P_{h,s,a}, P^o_{h,s,a})\right)\right].$$

The optimal robust value $V^*_{h,\lambda}$ satisfies the following robust dynamic programming procedure: Starting with $V^*_{H,\lambda} = 0$, doing backward iteration of $\mathcal{T}_v$, i.e., $V^*_{h,\lambda} = \mathcal{T}_v V^*_{h+1,\lambda}$, we get $V^*_{h,\lambda}$ for all $h \in [H]$. Furthermore, the robust regularized optimal policy π^* is a non-stationary deterministic Markov policy satisfying $\pi^*_h(\cdot) = \arg\max_a Q^*_{h,\lambda}(\cdot,a)$ for all $h \in [H]$ where

$$Q^*_{h,\lambda}(\cdot,a) = r_h(s,a) + \inf_{P_{h,s,a}\in\mathcal{P}_{h,s,a}} \left(\mathbb{E}_{s'\sim P_{h,s,a}}[V^*_{h+1}(s')] + \lambda D_\varphi(P_{h,s,a}, P^o_{h,s,a})\right).$$

Moreover, as $V^*_{H,\lambda} = 0 = Q^*_{H,\lambda}$, it suffices to backward iterate $\mathcal{T}$, i.e., do $Q^*_{h,\lambda} = \mathcal{T}Q^*_{h+1,\lambda}$ to get $Q^*_{h,\lambda}$ for all $h \in [H]$.

Proof. We start with the optimal robust value definition $V^*_{h,\lambda} = \max_\pi V^\pi_{h,\lambda} = \max_\pi \inf_{P\in\mathcal{P}} V^{h,\pi}_{P,r^\lambda_h}$. The value function claims in this statement are direct consequences of (Iyengar, 2005, Theorem 2.1 & 2.2) and (Zhang et al., 2024, Theorem 2) with the reward function r^λ_h.

It remains to prove Q^* dynamic programming with $\mathcal{T}$. That is, we establish $V^*_{h,\lambda}(\cdot) = \max_a Q^*_{h,\lambda}(\cdot,a)$ for all $h \in [H]$ with the dynamic programming of $\mathcal{T}$. We use induction to prove this. The base case is trivially true since $V^*_{H,\lambda} = 0 = Q^*_{H,\lambda}$. By $\mathcal{T}$, we have

$$Q^*_{h,\lambda}(s,a) = (\mathcal{T}Q^*_{h+1,\lambda})(s,a)$$

$$= r_h(s,a) + \inf_{P_{h,s,a} \in \mathcal{P}_{h,s,a}} \left(\mathbb{E}_{s' \sim P_{h,s,a}} [\max_{a'} Q^*_{h+1}(s',a')] + \lambda D_\varphi(P_{h,s,a}, P^o_{h,s,a}) \right)$$
$$= r_h(s,a) + \inf_{P_{h,s,a} \in \mathcal{P}_{h,s,a}} \left(\mathbb{E}_{s' \sim P_{h,s,a}} [V^*_{h+1}(s')] + \lambda D_\varphi(P_{h,s,a}, P^o_{h,s,a}) \right),$$

where the last equality follows by the induction hypothesis $V^*_{h+1,\lambda}(\cdot) = \max_a Q^*_{h+1,\lambda}(\cdot, a)$. Maximizing this both sides with action a and by the dynamic program $V^*_{h,\lambda} = \mathcal{T}_v V^*_{h+1,\lambda}$, we get $V^*_{h,\lambda}(\cdot) = \max_a Q^*_{h,\lambda}(\cdot, a)$. This completes the proof of this result. $\square$

E. Offline Robust φ-regularized RL Results

In this section, we set $V_{\max} = 1/(1-\gamma)$ whenever we use results from Proposition 3. In the following, we use constants c_1, c_2, c_3 from Proposition 3.

We first prove Proposition 1 that directly follows from Lemma 1.

Proof of Proposition 1. For each (s,a), consider the optimization problem in (3)

$$\begin{aligned}
\inf_{P_{s,a}\in\mathcal{P}_{s,a}} \left(\mathbb{E}_{s'\sim P_{s,a}}[V(s')] + \lambda D_\varphi(P_{s,a}, P^o_{s,a})\right) &= -\sup_{P_{s,a}\in\mathcal{P}_{s,a}} \left(\mathbb{E}_{s'\sim P_{s,a}}[-V(s')] - \lambda D_\varphi(P_{s,a}, P^o_{s,a})\right) \\
&\overset{(a)}{=} -\inf_{\eta'\in\mathbb{R}} (\lambda \mathbb{E}_{s'\sim P^o_{s,a}}[\varphi^*\left(\frac{-\eta' - V(s')}{\lambda}\right)] + \eta') \\
&\overset{(b)}{=} -\inf_{\eta\in\mathbb{R}} (\lambda \mathbb{E}_{s'\sim P^o_{s,a}}[\varphi^*\left(\frac{\eta - V(s')}{\lambda}\right)] - \eta) \\
&\overset{(c)}{=} -\inf_{\eta\in\Theta} (\lambda \mathbb{E}_{s'\sim P^o_{s,a}}[\varphi^*\left(\frac{\eta - V(s')}{\lambda}\right)] - \eta),
\end{aligned}$$

where (a) follows from Lemma 1, (b) by setting $\eta = -\eta'$, and (c) by Proposition 3. This completes the proof. □

We now prove Proposition 2 which mainly follows from Lemma 5.

Proof of Proposition 2. Since the conjugate function $\varphi^*(\cdot)$ is continuous, define a continuous function in η for each $(s,a)\in\mathcal{S}\times\mathcal{A}$ $h((s,a),\eta) = (\lambda \mathbb{E}_{s'\sim P^o_{s,a}} \varphi^*\left((\eta - \max_{a'} f(s',a'))/\lambda\right) - \eta)$. We observe $h((s,a),\eta)$ in $(s,a)\in\mathcal{S}\times\mathcal{A}$ is $\Sigma(\mathcal{S}\times\mathcal{A})$-measurable for each $\eta\in\Theta$, where Θ is a bounded real line. This lemma now directly follows by similar arguments in the proof of Panaganti et al. (2022, Lemma 1). □

Now we state a result and provide its proof for the empirical risk minimization on the dual parameter.

Proposition 7 (Dual Optimization Error Bound)**.** *Let $\widehat{g}_f$ be the dual optimization parameter from Algorithm 1 (Step 4) for the state-action value function f and let $\mathcal{T}_g$ be as defined in (7). With probability at least $1-\delta$, we have*

$$\sup_{f\in\mathcal{F}} \|\mathcal{T}f - \mathcal{T}_{\widehat{g}_f} f\|_{1,\mu} \le 2\gamma c_2 c_3 \sqrt{\frac{2\log(|\mathcal{G}|)}{N}} + 5c_1\sqrt{\frac{2\log(8|\mathcal{F}|/\delta)}{N}} + \gamma\varepsilon_{\mathcal{G}}.$$

Proof. We adapt the proof from Panaganti et al. (2022, Lemma 6). We first fix $f\in\mathcal{F}$. We will also invoke union bound for the supremum here. We recall from (8) that $\widehat{g}_f = \arg\min_{g\in\mathcal{G}} \widehat{L}_{\text{dual}}(g;f)$. From the robust Bellman equation, we directly obtain

$$\begin{aligned}
\|\mathcal{T}_{\widehat{g}_f} f - \mathcal{T}f\|_{1,\mu} &= \gamma(\mathbb{E}_{s,a\sim\mu}|\mathbb{E}_{s'\sim P^o_{s,a}}(\lambda\varphi^*((\widehat{g}_f(s,a) - \max_{a'} f(s',a'))/\lambda) - \widehat{g}_f(s,a)) \\
&\qquad - \inf_{\eta\in\Theta}(\lambda \mathbb{E}_{s'\sim P^o_{s,a}} \varphi^*\left((\eta - \max_{a'} f(s',a'))/\lambda\right) - \eta)|) \\
&\overset{(a)}{=} \gamma(\mathbb{E}_{s,a\sim\mu}\mathbb{E}_{s'\sim P^o_{s,a}}(\lambda\varphi^*((\widehat{g}_f(s,a) - \max_{a'} f(s',a'))/\lambda) - \widehat{g}_f(s,a)) \\
&\qquad - \mathbb{E}_{s,a\sim\mu}[\inf_{\eta\in\Theta}(\lambda \mathbb{E}_{s'\sim P^o_{s,a}} \varphi^*\left((\eta - \max_{a'} f(s',a'))/\lambda\right) - \eta))]) \\
&\overset{(b)}{=} \gamma(\mathbb{E}_{s,a\sim\mu,s'\sim P^o_{s,a}}(\lambda\varphi^*((\widehat{g}_f(s,a) - \max_{a'} f(s',a'))/\lambda) - \widehat{g}_f(s,a)) \\
&\qquad - \inf_{g\in L^1(\mu)} \mathbb{E}_{s,a\sim\mu,s'\sim P^o_{s,a}}(\lambda\varphi^*((g(s,a) - \max_{a'} f(s',a'))/\lambda) - g(s,a))) \\
&= \gamma(\mathbb{E}_{s,a\sim\mu,s'\sim P^o_{s,a}}(\lambda\varphi^*((\widehat{g}_f(s,a) - \max_{a'} f(s',a'))/\lambda) - \widehat{g}_f(s,a)) \\
&\qquad - \inf_{g\in\mathcal{G}} \mathbb{E}_{s,a\sim\mu,s'\sim P^o_{s,a}}(\lambda\varphi^*((g(s,a) - \max_{a'} f(s',a'))/\lambda) - g(s,a))) \\
&\qquad + \gamma(\inf_{g\in\mathcal{G}} \mathbb{E}_{s,a\sim\mu,s'\sim P^o_{s,a}}(\lambda\varphi^*((g(s,a) - \max_{a'} f(s',a'))/\lambda) - g(s,a))
\end{aligned}$$

$$
\begin{aligned}
&\quad - \inf_{g \in L^1(\mu)} \mathbb{E}_{s,a\sim\mu,s'\sim P^o_{s,a}}(\lambda\varphi^*((g(s,a) - \max_{a'} f(s',a'))/\lambda) - g(s,a))) \\
&\overset{(c)}{\leq} \gamma(\mathbb{E}_{s,a\sim\mu,s'\sim P^o_{s,a}}(\lambda\varphi^*((\widehat{g}_f(s,a) - \max_{a'} f(s',a'))/\lambda) - \widehat{g}_f(s,a)) \\
&\qquad - \inf_{g\in\mathcal{G}} \mathbb{E}_{s,a\sim\mu,s'\sim P^o_{s,a}}(\lambda\varphi^*((g(s,a) - \max_{a'} f(s',a'))/\lambda) - g(s,a))) + \gamma\varepsilon_{\mathcal{G}} \\
&\overset{(d)}{\leq} 2\gamma c_2 c_3\sqrt{\frac{2\log(|\mathcal{G}|)}{N}} + 5c_1\sqrt{\frac{2\log(8/\delta)}{N}} + \gamma\varepsilon_{\mathcal{G}}.
\end{aligned}
$$

(a) follows since $\inf_g h(g) \leq h(\widehat{g}_f)$. (b) follows from Proposition 2. (c) follows from the approximate dual realizability assumption (Assumption 3).

For (d), we consider the loss function $l(g,(s,a,s')) = \lambda\varphi^*((g(s,a) - \max_{a'} f(s',a'))/\lambda) - g(s,a)$ (for e.g. $l(g,(s,a,s')) = [(g(s,a) + 2\lambda - \max_{a'} f(s',a'))_+^2]/4\lambda - \lambda - g(s,a)$) and dataset $\mathcal{D} = \{s_i, a_i, s_i'\}_{i=1}^N$. Since $f \in \mathcal{F}$ and $g \in \mathcal{G}$, we note that $|l(g,(s,a,s'))| \leq c_1$, where the value of $c_1 > 0$ depend on specific forms of φ^* as demonstrated in Proposition 3. Furthermore, take $l(g,(s,a,s'))$ to be c_2-Lipschitz in g and $|g(s,a)| \leq c_3$, since $g \in \mathcal{G}$, for some positive constants c_2 and c_3. Again, these constants depend on specific forms of φ^* as demonstrated in Proposition 3. With these insights, we can apply the empirical risk minimization result in Lemma 4 to get (d).

With union bound, with probability at least $1-\delta$, we finally get

$$
\sup_{f\in\mathcal{F}} \|\mathcal{T}f - \mathcal{T}_{\widehat{g}_f} f\|_{1,\mu} \leq 2\gamma c_2 c_3\sqrt{\frac{2\log(|\mathcal{G}|)}{N}} + 5c_1\sqrt{\frac{2\log(8|\mathcal{F}|/\delta)}{N}} + \gamma\varepsilon_{\mathcal{G}},
$$

which concludes the proof. □

We next prove the least-squares generalization bound for the RFQI algorithm.

Proposition 8 (Least squares generalization bound)**.** *Let $\widehat{f}_g$ be the least-squares solution from Algorithm 1 (Step 5) for the state-action value function f and dual variable function g. Let $\mathcal{T}_g$ be as defined in (7). Then, with probability at least $1-\delta$, we have*

$$
\sup_{f\in\mathcal{F}}\sup_{g\in\mathcal{G}} \|\mathcal{T}_g f - \widehat{f}_g\|_{2,\mu} \leq \sqrt{6\varepsilon_{\mathcal{F}}} + \sqrt{\frac{2}{(1-\gamma)^2} + 18(1+\gamma c_1)}\sqrt{\frac{18\log(2|\mathcal{F}||\mathcal{G}|/\delta)}{N}}.
$$

Proof. We adapt the least-squares generalization bound given in Agarwal et al. (2019, Lemma A.11) to our setting. We recall from (9) that $\widehat{f}_g = \arg\min_{Q\in\mathcal{F}} \widehat{L}_{\mathrm{robQ}}(Q; f, g)$. We first fix functions $f \in \mathcal{F}$ and $g \in \mathcal{G}$. For any function $f' \in \mathcal{F}$, we define random variables $z_i^{f'}$ as

$$
z_i^{f'} = (f'(s_i,a_i) - y_i)^2 - ((\mathcal{T}_g f)(s_i,a_i) - y_i)^2,
$$

where $y_i = r_i - \gamma\lambda\varphi^*((g(s_i,a_i) - \max_{a'} f(s_i',a'))/\lambda) + \gamma g(s_i,a_i)$, and $(s_i,a_i,s_i') \in \mathcal{D}$ with $(s_i,a_i) \sim \mu$, $s_i' \sim P^o_{s_i,a_i}$. It is straightforward to note that for a given (s_i,a_i), we have $\mathbb{E}_{s_i'\sim P^o_{s_i,a_i}}[y_i] = (\mathcal{T}_g f)(s_i,a_i)$. We note the randomness of $z_i^{f'}$ given $f, f' \in \mathcal{F}$ and $g \in \mathcal{G}$ is from the dataset pairs (s_i,a_i,s_i').

Since $f, f' \in \mathcal{F}$ and $g \in \mathcal{G}$, from Proposition 3, we write both $(\mathcal{T}_g f)(s_i,a_i), y_i \leq 1 + \gamma c_1$, where the value of $c_1 > 0$ depend on specific forms of φ^*. Using this, we obtain the first moment and an upper-bound for the second moment of $z_i^{f'}$ as follows:

$$
\begin{aligned}
\mathbb{E}_{s_i'\sim P^o_{s_i,a_i}}[z_i^{f'}] &= \mathbb{E}_{s_i'\sim P^o_{s_i,a_i}}[(f'(s_i,a_i) - (\mathcal{T}_g f)(s_i,a_i)) \cdot (f'(s_i,a_i) + (\mathcal{T}_g f)(s_i,a_i) - 2y_i)] \\
&= (f'(s_i,a_i) - (\mathcal{T}_g f)(s_i,a_i))^2, \\
\mathbb{E}_{s_i'\sim P^o_{s_i,a_i}}[(z_i^{f'})^2] &= \mathbb{E}_{s_i'\sim P^o_{s_i,a_i}}[(f'(s_i,a_i) - (\mathcal{T}_g f)(s_i,a_i))^2 \cdot (f'(s_i,a_i) + (\mathcal{T}_g f)(s_i,a_i) - 2y_i)^2] \\
&= (f'(s_i,a_i) - (\mathcal{T}_g f)(s_i,a_i))^2 \cdot \mathbb{E}_{s_i'\sim P^o_{s_i,a_i}}[(f'(s_i,a_i) + (\mathcal{T}_g f)(s_i,a_i) - 2y_i)^2] \\
&\leq C_1(f'(s_i,a_i) - (\mathcal{T}_g f)(s_i,a_i))^2,
\end{aligned}
$$

where $C_1 = \frac{2}{(1-\gamma)^2} + 18(1+\gamma c_1)$. This immediately implies that

$$\mathbb{E}_{s_i,a_i\sim\mu,s_i'\sim P^o_{s_i,a_i}}[z_i^{f'}] = \|\mathcal{T}_g f - f'\|_{2,\mu}^2,$$
$$\mathbb{E}_{s_i,a_i\sim\mu,s_i'\sim P^o_{s_i,a_i}}[(z_i^{f'})^2] \le C_1 \|\mathcal{T}_g f - f'\|_{2,\mu}^2.$$

From these calculations, it is also straightforward to see that $|z_i^{f'} - \mathbb{E}_{s_i,a_i\sim\mu,s_i'\sim P^o_{s_i,a_i}}[z_i^{f'}]| \le 2C_1$ almost surely.

Now, using the Bernstein's inequality (Lemma 2), together with a union bound over all $f' \in \mathcal{F}$, with probability at least $1-\delta$, we have

$$|\|\mathcal{T}_g f - f'\|_{2,\mu}^2 - \frac{1}{N}\sum_{i=1}^N z_i^{f'}| \le \sqrt{\frac{2C_1\|\mathcal{T}_g f - f'\|_{2,\mu}^2 \log(2|\mathcal{F}|/\delta)}{N}} + \frac{2C_1\log(2|\mathcal{F}|/\delta)}{3N}, \tag{23}$$

for all $f' \in \mathcal{F}$. This expression coincides with Panaganti et al. (2022, Eq.(15)). Thus, following the proof of Panaganti et al. (2022, Lemma 7), we finally get

$$\|\mathcal{T}_g f - \widehat{f}_g\|_{2,\mu}^2 \le 6\varepsilon_{\mathcal{F}} + \frac{9C_1\log(4|\mathcal{F}|/\delta)}{N}. \tag{24}$$

We note a fact $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$. Now, using union bound for $f \in \mathcal{F}$ and $g \in \mathcal{G}$, with probability at least $1-\delta$, we finally obtain

$$\sup_{f\in\mathcal{F}}\sup_{g\in\mathcal{G}} \|\mathcal{T}_g f - \widehat{f}_g\|_{2,\mu} \le \sqrt{6\varepsilon_{\mathcal{F}}} + \sqrt{\frac{18C_1\log(2|\mathcal{F}||\mathcal{G}|/\delta)}{N}}.$$

This completes the least-squares generalization bound analysis for the robust regularized Bellman updates. □

We are now ready to prove the main theorem.

E.1. Proof of Theorem 1

Theorem 3 (Restatement of Theorem 1)**.** *Let Assumptions 1 to 3 hold. Let π_K be the RPQ algorithm policy after K iterations. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$V^{\pi^*} - V^{\pi_K} \le \frac{2\gamma^K}{(1-\gamma)^2} + \frac{2\sqrt{C}}{(1-\gamma)^2}\Big(2\gamma c_2 c_3\sqrt{\frac{2\log(|\mathcal{G}|)}{N}} + 5c_1\sqrt{\frac{2\log(8|\mathcal{F}|/\delta)}{N}} + \gamma\varepsilon_{\mathcal{G}}\Big)$$
$$+ \frac{2\sqrt{C}}{(1-\gamma)^2}\Big(\sqrt{6\varepsilon_{\mathcal{F}}} + \sqrt{\frac{2}{(1-\gamma)^2} + 18(1+\gamma c_1)}\sqrt{\frac{18\log(2|\mathcal{F}||\mathcal{G}|/\delta)}{N}}\Big).$$

Proof. We let $V_k(s) = Q_k(s,\pi_k(s))$ for every $s \in \mathcal{S}$. Since π_k is the greedy policy w.r.t Q_k, we also have $V_k(s) = Q_k(s,\pi_k(s)) = \max_a Q_k(s,a)$. We recall that $V^* = V^{\pi^*}$ and $Q^* = Q^{\pi^*}$. We also recall from Section 2 that Q^{π^*} is a fixed-point of the robust Bellman operator $\mathcal{T}$ defined in (3). We also note that the same holds true for any stationary deterministic policy π from Yang et al. (2023) that Q^π satisfies $Q^\pi(s,a) = r(s,a) + \gamma\min_{P_{s,a}\ll P^o_{s,a}}(\mathbb{E}_{s'\sim P_{s,a}}[V^\pi(s')] + \lambda D_\varphi(P_{s,a}, P^o_{s,a}))$. We now adapt the proof of Panaganti et al. (2022, Theorem 1) using the RRBE in its primal form (3) directly instead of its dual form (4).

We first characterize the performance decomposition between V^{π^*} and V^{π_K}. We recall the initial state distribution d_0. Since $V^{\pi^*}(s) \ge V^{\pi_K}(s)$ for any $s \in \mathcal{S}$, we observe that

$$\begin{aligned}
0 \le& \mathbb{E}_{s_0\sim d_0}[V^{\pi^*}(s_0) - V^{\pi_K}(s_0)] = \mathbb{E}_{s_0\sim d_0}[(V^{\pi^*}(s_0) - V_K(s_0)) - (V^{\pi_K}(s_0) - V_K(s_0))] \\
&= \mathbb{E}_{s_0\sim d_0}[(Q^{\pi^*}(s_0,\pi^*(s_0)) - Q_K(s_0,\pi_K(s_0))) - (Q^{\pi_K}(s_0,\pi_K(s_0)) - Q_K(s_0,\pi_K(s_0)))] \\
&\overset{(a)}{\le} \mathbb{E}_{s_0\sim d_0}[Q^{\pi^*}(s_0,\pi^*(s_0)) - Q_K(s_0,\pi^*(s_0)) + Q_K(s_0,\pi_K(s_0)) - Q^{\pi_K}(s_0,\pi_K(s_0))]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{s_0\sim d_0}[Q^{\pi^*}(s_0,\pi^*(s_0)) - Q_K(s_0,\pi^*(s_0)) + Q_K(s_0,\pi_K(s_0)) - Q^{\pi^*}(s_0,\pi_K(s_0)) \\
&\qquad + Q^{\pi^*}(s_0,\pi_K(s_0)) - Q^{\pi_K}(s_0,\pi_K(s_0))] \\
&\overset{(b)}{\le} \mathbb{E}_{s_0\sim d_0}[Q^{\pi^*}(s_0,\pi^*(s_0)) - Q_K(s_0,\pi^*(s_0)) + Q_K(s_0,\pi_K(s_0)) - Q^{\pi^*}(s_0,\pi_K(s_0)) \\
&\qquad + \gamma[\min_{P_{s_0,\pi_K(s_0)}\ll P^o_{s_0,\pi_K(s_0)}} (\mathbb{E}_{s_1\sim P_{s_0,\pi_K(s_0)}}[V^{\pi^*}(s_1)] + \lambda D_\varphi(P_{s_0,\pi_K(s_0)}, P^o_{s_0,\pi_K(s_0)})) \\
&\qquad - \min_{P_{s_0,\pi_K(s_0)}\ll P^o_{s_0,\pi_K(s_0)}} (\mathbb{E}_{s_1\sim P_{s_0,\pi_K(s_0)}}[V^{\pi_K}(s_1)] + \lambda D_\varphi(P_{s_0,\pi_K(s_0)}, P^o_{s_0,\pi_K(s_0)}))]] \\
&\overset{(c)}{\le} \mathbb{E}_{s_0\sim d_0}[|Q^{\pi^*}(s_0,\pi^*(s_0)) - Q_K(s_0,\pi^*(s_0))|] + \mathbb{E}_{s_0\sim d_0}[|Q^{\pi^*}(s_0,\pi_K(s_0)) - Q_K(s_0,\pi_K(s_0))|] \\
&\qquad + \gamma\mathbb{E}_{s_0\sim d_0}\mathbb{E}_{s_1\sim P^{\pi_K,\min}_{s_0,\pi_K(s_0)}}(|V^{\pi^*}(s_1) - V^{\pi_K}(s_1)|) \\
&\overset{(d)}{\le} \sum_{h=0}^{\infty}\gamma^h\cdot\left(\mathbb{E}_{s\sim d_{h,\pi_K}}[|Q^{\pi^*}(s,\pi^*(s)) - Q_K(s,\pi^*(s))| + |Q^{\pi^*}(s,\pi_K(s)) - Q_K(s,\pi_K(s))|]\right),
\end{aligned} \tag{25}
$$

where (a) follows from the fact that π_K is the greedy policy with respect to Q_K, (b) from the Bellman equations, and (c) from the following definition

$$
P^{\pi_K,\min}_{s,\pi_K(s)} \in \operatorname*{arg\,min}_{P_{s,\pi_K(s)}\ll P^o_{s,\pi_K(s)}} (\mathbb{E}_{s'\sim P_{s,\pi_K(s)}}[V^{\pi_K}(s')] + \lambda D_\varphi(P_{s,\pi_K(s)}, P^o_{s,\pi_K(s)})).
$$

We note that this worse-case model distribution can be non-unique and we just pick one by an arbitrary deterministic rule. We emphasize that this model distribution is used only in analysis which is not required in the algorithm. Finally, (d) follows with telescoping over $|V^{\pi^*} - V^{\pi_K}|$ by defining a state distribution $d_{h,\pi_K} \in \Delta(\mathcal{S})$, for all natural numbers $h \ge 0$, as

$$
d_{h,\pi_K} = \begin{cases} d_0 & \text{if } h = 0, \\ P^{\pi_K,\min}_{s',\pi_K(s')} & \text{otherwise, with } s' \sim d_{h-1,\pi_K}. \end{cases}
$$

We note that such state distribution proof ideas are commonly used in the offline RL literature (Agarwal et al., 2019; Panaganti et al., 2022; Bruns-Smith & Zhou, 2023; Zhang et al., 2024).

For (25), with the ν-norm notation i.e. $\|f\|^2_{p,\nu} = (\mathbb{E}_{s,a\sim\nu}|f(s,a)|^p)^{1/p}$ for any $\nu \in \Delta(\mathcal{S}\times\mathcal{A})$, we have

$$
\mathbb{E}_{s_0\sim d_0}[V^{\pi^*}] - \mathbb{E}_{s_0\sim d_0}[V^{\pi_K}] \le \sum_{h=0}^{\infty}\gamma^h\left(\|Q^{\pi^*} - Q_K\|_{1,d_{h,\pi_K}\circ\pi^*} + \|Q^{\pi^*} - Q_K\|_{1,d_{h,\pi_K}\circ\pi_K}\right), \tag{26}
$$

where the state-action distributions are $d_{h,\pi_K}\circ\pi^*(s,a) \propto d_{h,\pi_K}(s)\mathbb{1}(a = \pi^*(s))$ and $d_{h,\pi_K}\circ\pi_K(s,a) \propto d_{h,\pi_K}(s)\mathbb{1}(a = \pi_K(s))$. We now analyze the above two terms treating either $d_{h,\pi_K}\circ\pi^*$ or $d_{h,\pi_K}\circ\pi_K$ as a state-action distribution ν satisfying Assumption 1. First, considering any $s,a\sim\nu$ satisfying $Q^{\pi^*}(s,a) \ge Q_K(s,a)$ we have

$$
\begin{aligned}
0 \le& Q^{\pi^*}(s,a) - Q_K(s,a) \le Q^{\pi^*}(s,a) - \mathcal{T}Q_{K-1}(s,a) + |\mathcal{T}Q_{K-1}(s,a) - Q_K(s,a)| \\
&\le Q^{\pi^*}(s,a) - \mathcal{T}Q_{K-1}(s,a) + \|\mathcal{T}Q_{K-1} - Q_K\|_{1,\nu} \\
&\overset{(e)}{\le} Q^{\pi^*}(s,a) - \mathcal{T}Q_{K-1}(s,a) + \sqrt{C}\|\mathcal{T}Q_{K-1} - Q_K\|_{1,\mu} \\
&\overset{(f)}{=} \gamma[\min_{P_{s,a}\ll P^o_{s,a}} (\mathbb{E}_{s'\sim P_{s,a}}[\max_{a'} Q^{\pi^*}(s',a')] + \lambda D_\varphi(P_{s,a}, P^o_{s,a})) \\
&\qquad - \min_{P_{s,a}\ll P^o_{s,a}} (\mathbb{E}_{s'\sim P_{s,a}}[\max_{a'} Q_{K-1}(s',a')] + \lambda D_\varphi(P_{s,a}, P^o_{s,a}))] \\
&\qquad + \sqrt{C}\|\mathcal{T}Q_{K-1} - Q_K\|_{1,\mu} \\
&\overset{(g)}{\le} \gamma(\mathbb{E}_{s'\sim P^{Q_{K-1},\min}_{s,a}}(\max_{a'} Q^{\pi^*}(s',a') - \max_{a'} Q_{K-1}(s',a'))) + \sqrt{C}\|\mathcal{T}Q_{K-1} - Q_K\|_{1,\mu} \\
&\overset{(h)}{\le} \gamma(\mathbb{E}_{s'\sim P^{Q_{K-1},\min}_{s,a}} \max_{a'} |Q^{\pi^*}(s',a') - Q_{K-1}(s',a')|) + \sqrt{C}\|\mathcal{T}Q_{K-1} - Q_K\|_{1,\mu},
\end{aligned} \tag{27}
$$

where (e) follows by the concentrability assumption (Assumption 1), (f) from Bellman equation, operator $\mathcal{T}$, (g) follows, similarly as step (c), from the following definition

$$P_{s,a}^{Q_{K-1},\min} \in \underset{P_{s,a} \ll P_{s,a}^o}{\arg\min} \left(\mathbb{E}_{s' \sim P_{s,a}}[\max_{a'} Q_{K-1}(s',a')] + \lambda D_\varphi(P_{s,a}, P_{s,a}^o)\right).$$

We again emphasize that this model distribution is analysis-specific and we just pick one by an arbitrary deterministic rule since it may not be unique. (h) follows by the fact $|\sup_x p(x) - \sup_x q(x)| \leq \sup_x |p(x) - q(x)|$. Now, by replacing $P^{Q_{K-1},\min}$ with $P^{Q^{\pi^*},\min}$ in step (g) and repeating the steps for any $s, a \sim \nu$ satisfying $Q^{\pi^*}(s,a) \leq Q_K(s,a)$, we get

$$0 \leq Q_K(s,a) - Q^{\pi^*}(s,a) \leq \gamma(\mathbb{E}_{s' \sim P_{s,a}^{Q^{\pi^*},\min}} \max_{a'} |Q^{\pi^*}(s',a') - Q_{K-1}(s',a')|) + \sqrt{C}\|\mathcal{T}Q_{K-1} - Q_K\|_{1,\mu}. \quad (28)$$

We immediately note that both $P_{s,a}^{Q_{K-1},\min}$ and $P_{s,a}^{Q^{\pi^*},\min}$ satisfies $D_\varphi(P_{s,a}^{Q_{K-1},\min}, P_{s,a}^o) \leq 1/(\lambda(1-\gamma))$ and $D_\varphi(P_{s,a}^{Q^{\pi^*},\min}, P_{s,a}^o) \leq 1/(\lambda(1-\gamma))$, which follows by their definition and the facts $Q_{K-1} \in \mathcal{F}$, $\|Q^{\pi^*}\|_\infty \leq 1/(1-\gamma)$. Define the state-action probability distribution ν' as, for any s', a',

$$\begin{aligned}\nu'(s',a') = &\sum_{s,a} \nu(s,a)\mathbb{1}\{Q^{\pi^*}(s,a) > Q_K(s,a)\}P_{s,a}^{Q_{K-1},\min}(s')\mathbb{1}\{a' = \arg\max_b |Q^{\pi^*}(s',b) - Q_{K-1}(s',b)|\} \\ &+ \sum_{s,a} \nu(s,a)\mathbb{1}\{Q^{\pi^*}(s,a) \leq Q_K(s,a)\}P_{s,a}^{Q^{\pi^*},\min}(s')\mathbb{1}\{a' = \arg\max_b |Q^{\pi^*}(s',b) - Q_{K-1}(s',b)|\}.\end{aligned}$$

Now, we can combine (27)-(28) as follows

$$\begin{aligned}\|Q^{\pi^*} - Q_K\|_{1,\nu} &\leq \gamma\|Q^{\pi^*} - Q_{K-1}\|_{1,\nu'} + \sqrt{C}\|\mathcal{T}Q_{K-1} - Q_K\|_{1,\mu} \\ &\overset{(i)}{\leq} \gamma\|Q^{\pi^*} - Q_{K-1}\|_{1,\nu'} + \sqrt{C}\|\mathcal{T}_{g_{K-1}}Q_{K-1} - Q_K\|_{2,\mu} + \sqrt{C}\|\mathcal{T}Q_{K-1} - \mathcal{T}_{g_{K-1}}Q_{K-1}\|_{1,\mu},\end{aligned}$$

where (i) uses the fact $\|\cdot\|_{1,\mu} \leq \|\cdot\|_{2,\mu}$.

Now, by recursion until iteration 0, we get

$$\begin{aligned}\|Q^{\pi^*} - Q_K\|_{1,\nu} &\leq \gamma^K \sup_{\bar{\nu}} \|Q^{\pi^*} - Q_0\|_{1,\bar{\nu}} + \sqrt{C}\sum_{t=0}^{K-1}\gamma^t\|\mathcal{T}Q_{K-1-t} - \mathcal{T}_{g_{K-1-t}}Q_{K-1-t}\|_{1,\mu} \\ &\quad + \sqrt{C}\sum_{t=0}^{K-1}\gamma^t\|\mathcal{T}_{g_{K-1-t}}Q_{K-1-t} - Q_{K-t}\|_{2,\mu} \\ &\overset{(j)}{\leq} \frac{\gamma^K}{1-\gamma} + \sqrt{C}\sum_{t=0}^{K-1}\gamma^t\|\mathcal{T}Q_{K-1-t} - \mathcal{T}_{g_{K-1-t}}Q_{K-1-t}\|_{1,\mu} \\ &\quad + \sqrt{C}\sum_{t=0}^{K-1}\gamma^t\|\mathcal{T}_{g_{K-1-t}}Q_{K-1-t} - Q_{K-t}\|_{2,\mu} \\ &\overset{(k)}{\leq} \frac{\gamma^K}{1-\gamma} + \frac{\sqrt{C}}{1-\gamma}\sup_{f\in\mathcal{F}}\|\mathcal{T}f - \mathcal{T}_{\widehat{g}_f}f\|_{1,\mu} + \frac{\sqrt{C}}{1-\gamma}\sup_{f\in\mathcal{F}}\|\mathcal{T}_{\widehat{g}_f}f - \widehat{f}_{\widehat{g}_f}\|_{2,\mu} \\ &\leq \frac{\gamma^K}{1-\gamma} + \frac{\sqrt{C}}{1-\gamma}\sup_{f\in\mathcal{F}}\|\mathcal{T}f - \mathcal{T}_{\widehat{g}_f}f\|_{1,\mu} + \frac{\sqrt{C}}{1-\gamma}\sup_{f\in\mathcal{F}}\sup_{g\in\mathcal{G}}\|\mathcal{T}_g f - \widehat{f}_g\|_{2,\mu}. \quad (29)\end{aligned}$$

where (j) follows since $|Q^{\pi^*}(s,a)| \leq 1/(1-\gamma), Q_0(s,a) = 0$, and (k) follows since $\widehat{g}_f$ is the dual variable function from the algorithm for the state-action value function f and $\widehat{f}_g$ as the least squares solution from the algorithm for the state-action value function f and dual variable function g pair.

Now, using Lemma 7 and Lemma 8 to bound (29), and then combining it with (26), completes the proof of this theorem. □

E.2. Specialized Result for TV φ-divergence

We now state and prove the improved (in terms of assumptions) result for TV φ-divergence.

Assumption 9 (Concentrability)**.** *There exists a finite constant $C_{\text{tv}} > 0$ such that for any $\nu \in \{d_{\pi,P^o}\} \subseteq \Delta(\mathcal{S} \times \mathcal{A})$ for any policy π (can be non-stationary as well), we have $\|\nu/\mu\|_\infty \leq \sqrt{C_{\text{tv}}}$.*

Assumption 10 (Fail-state)**.** *There is a fail state s_f such that $r(s_f, a) = 0$ and $P_{s_f,a}(s_f) = 1$, for all $a \in \mathcal{A}$ and $P \in \mathcal{P}$ satisfying $D_{\text{TV}}(P_{s',a'}, P^o_{s',a'}) \leq \max\{1, 1/(\lambda(1-\gamma))\}$ for all s', a'.*

Theorem 4. *Let Assumptions 2, 3, 9 and 10 hold. Let π_K be the RPQ algorithm policy after K iterations. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$
\begin{aligned}
V^{\pi^*} - V^{\pi_K} \leq & \frac{2\gamma^K}{(1-\gamma)^2} + \frac{2\sqrt{C_{\text{tv}}}}{(1-\gamma)^2}(2\gamma c_2 c_3\sqrt{\frac{2\log(|\mathcal{G}|)}{N}} + 5c_1\sqrt{\frac{2\log(8|\mathcal{F}|/\delta)}{N}} + \gamma\varepsilon_{\mathcal{G}}) \\
& + \frac{2\sqrt{C_{\text{tv}}}}{(1-\gamma)^2}(\sqrt{6\varepsilon_{\mathcal{F}}} + \sqrt{\frac{2}{(1-\gamma)^2} + 18(1+\gamma c_1)}\sqrt{\frac{18\log(2|\mathcal{F}||\mathcal{G}|/\delta)}{N}}),
\end{aligned}
$$

with $c_1 = 2\lambda + (1/(1-\gamma)), c_2 = 2, c_3 = \lambda/2$.

Proof. We can now further use the dual form (4) under Assumption 10. We again start by characterizing the performance decomposition between V^{π^*} and V^{π_K}. This proof largely follows the proofs of Theorem 1 and Panaganti et al. (2022, Theorem 1). In particular, we use the total variation RRBE its dual form (4) under Assumption 10 in this proof. That is, for all π and $Q \in \mathcal{F}$, from (17) we have

$$
\begin{aligned}
Q^\pi(s,a) &= r(s,a) - \inf_{\eta \in [0,\lambda]} \left(\mathbb{E}_{s' \sim P^o_{s,a}}[(\eta - V^\pi(s'))_+] - \eta\right) \text{ and} \\
(\mathcal{T}Q)(s,a) &= r(s,a) - \inf_{\eta \in [0,\lambda]} \left(\mathbb{E}_{s' \sim P^o_{s,a}}[(\eta - \max_{a'} Q(s',a'))_+] - \eta\right).
\end{aligned} \tag{30}
$$

We recall the initial state distribution d_0. Since $V^{\pi^*}(s) \geq V^{\pi_K}(s)$ for any $s \in \mathcal{S}$, we begin with step (b) in Theorem 1:

$$
\begin{aligned}
0 \leq & \mathbb{E}_{s_0 \sim d_0}[V^{\pi^*}(s_0) - V^{\pi_K}(s_0)] \\
\leq & \mathbb{E}_{s_0 \sim d_0}[Q^{\pi^*}(s_0, \pi^*(s_0)) - Q_K(s_0, \pi^*(s_0)) + Q_K(s_0, \pi_K(s_0)) - Q^{\pi^*}(s_0, \pi_K(s_0)) \\
& + \gamma[\min_{P_{s_0,\pi_K(s_0)} \ll P^o_{s_0,\pi_K(s_0)}} (\mathbb{E}_{s_1 \sim P_{s_0,\pi_K(s_0)}}[V^{\pi^*}(s_1)] + \lambda D_\varphi(P_{s_0,\pi_K(s_0)}, P^o_{s_0,\pi_K(s_0)})) \\
& \quad - \min_{P_{s_0,\pi_K(s_0)} \ll P^o_{s_0,\pi_K(s_0)}} (\mathbb{E}_{s_1 \sim P_{s_0,\pi_K(s_0)}}[V^{\pi_K}(s_1)] + \lambda D_\varphi(P_{s_0,\pi_K(s_0)}, P^o_{s_0,\pi_K(s_0)}))]] \\
\overset{(a)}{\leq} & \mathbb{E}_{s_0 \sim d_0}[|Q^{\pi^*}(s_0, \pi^*(s_0)) - Q_K(s_0, \pi^*(s_0))|] + \mathbb{E}_{s_0 \sim d_0}[|Q^{\pi^*}(s_0, \pi_K(s_0)) - Q_K(s_0, \pi_K(s_0))|] \\
& + \gamma \mathbb{E}_{s_0 \sim d_0} \sup_\eta (\mathbb{E}_{s_1 \sim P^o_{s_0,\pi_K(s_0)}}((\eta - V^{\pi_K}(s_1))_+ - (\eta - V^{\pi^*}(s_1))_+)) \\
\overset{(b)}{\leq} & \mathbb{E}_{s_0 \sim d_0}[|Q^{\pi^*}(s_0, \pi^*(s_0)) - Q_K(s_0, \pi^*(s_0))|] + \mathbb{E}_{s_0 \sim d_0}[|Q^{\pi^*}(s_0, \pi_K(s_0)) - Q_K(s_0, \pi_K(s_0))|] \\
& + \gamma \mathbb{E}_{s_0 \sim d_0} \mathbb{E}_{s_1 \sim P^o_{s_0,\pi_K(s_0)}}(|V^{\pi^*}(s_1) - V^{\pi_K}(s_1)|) \\
\overset{(c)}{\leq} & \sum_{h=0}^{\infty} \gamma^h \times \left(\mathbb{E}_{s \sim d_{h,\pi_K}}[|Q^{\pi^*}(s, \pi^*(s)) - Q_K(s, \pi^*(s))| + |Q^{\pi^*}(s, \pi_K(s)) - Q_K(s, \pi_K(s))|]\right),
\end{aligned} \tag{31}
$$

where (a) follows from (30) and the fact $|\sup_x f(x) - \sup_x g(x)| \leq \sup_x |f(x) - g(x)|$, (b) follows from the facts $(x)_+ - (y)_+ \leq (x-y)_+$ and $(x)_+ \leq |x|$ for any $x, y \in \mathbb{R}$. We make an important note here in step (b) regarding the dependence on the nominal model P^o distribution unlike in step (c) in the proof of Theorem 1. This important step helps us improve the concentrability assumption in further analysis. Finally, (c) follows with telescoping over $|V^{\pi^*} - V^{\pi_K}|$ by defining a new state distribution $d_{h,\pi_K} \in \Delta(\mathcal{S})$, for all natural numbers $h \geq 0$, as

$$
d_{h,\pi_K} = \begin{cases} d_0 & \text{if } h = 0, \\ P^o_{s',\pi_K(s')} & \text{otherwise, with } s' \sim d_{h-1,\pi_K}. \end{cases}
$$

For (31), with the ν-norm notation i.e. $\|f\|_{p,\nu}^2 = (\mathbb{E}_{s,a\sim\nu}|f(s,a)|^p)^{1/p}$ for any $\nu \in \Delta(\mathcal{S}\times\mathcal{A})$, we have

$$\begin{aligned}\mathbb{E}_{s_0\sim d_0}[V^{\pi^*}] - \mathbb{E}_{s_0\sim d_0}[V^{\pi_K}] &\leq \sum_{h=0}^{\infty}\gamma^h\Big(\|Q^{\pi^*} - Q_K\|_{1,d_{h,\pi_K}\circ\pi^*} + \|Q^{\pi^*} - Q_K\|_{1,d_{h,\pi_K}\circ\pi_K}\Big),\\ &\leq \sum_{h=0}^{\infty}\gamma^h(2\sup_{\nu}\|Q^{\pi^*} - Q_K\|_{1,\nu}),\end{aligned} \tag{32}$$

where the second inequality follows since both $d_{h,\pi_K}\circ\pi^*$ and $d_{h,\pi_K}\circ\pi_K$ satisfy Assumption 9. We now analyze the summand in (26):

$$\begin{aligned}
&\|Q^{\pi^*} - Q_K\|_{1,\nu} \leq \|Q^{\pi^*} - \mathcal{T}Q_{K-1}\|_{1,\nu} + \|\mathcal{T}Q_{K-1} - Q_K\|_{1,\nu}\\
&\overset{(d)}{\leq} \|Q^{\pi^*} - \mathcal{T}Q_{K-1}\|_{1,\nu} + \sqrt{C_{\mathrm{tv}}}\|\mathcal{T}Q_{K-1} - Q_K\|_{1,\mu}\\
&= (\mathbb{E}_{s,a\sim\nu}|Q^{\pi^*}(s,a) - \mathcal{T}Q_{K-1}(s,a)|) + \sqrt{C_{\mathrm{tv}}}\|\mathcal{T}Q_{K-1} - Q_K\|_{1,\mu}\\
&\overset{(e)}{\leq} (\mathbb{E}_{s,a\sim\nu}\gamma\sup_{\eta}|\mathbb{E}_{s'\sim P^o_{s,a}}((\eta - \max_{a'}Q_{K-1}(s',a'))_+ - (\eta - \max_{a'}Q^{\pi^*}(s',a'))_+)|)\\
&\qquad\qquad + \sqrt{C_{\mathrm{tv}}}\|\mathcal{T}Q_{K-1} - Q_K\|_{1,\mu}\\
&\overset{(f)}{\leq} (\mathbb{E}_{s,a\sim\nu}|\mathbb{E}_{s'\sim P^o_{s,a}}(\max_{a'}Q^{\pi^*}(s',a') - \max_{a'}Q_{K-1}(s',a'))_+|) + \sqrt{C_{\mathrm{tv}}}\|\mathcal{T}Q_{K-1} - Q_K\|_{1,\mu}\\
&\overset{(g)}{\leq} \gamma(\mathbb{E}_{s,a\sim\nu}\mathbb{E}_{s'\sim P^o_{s,a}}\max_{a'}|Q^{\pi^*}(s',a') - Q_{K-1}(s',a')|) + \sqrt{C_{\mathrm{tv}}}\|\mathcal{T}Q_{K-1} - Q_K\|_{1,\mu}\\
&\overset{(h)}{\leq} \gamma\|Q^{\pi^*} - Q_{K-1}\|_{1,\nu'} + \sqrt{C_{\mathrm{tv}}}\|\mathcal{T}Q_{K-1} - Q_K\|_{1,\mu}\\
&\overset{(i)}{\leq} \gamma\|Q^{\pi^*} - Q_{K-1}\|_{1,\nu'} + \sqrt{C_{\mathrm{tv}}}\|\mathcal{T}_{g_{K-1}}Q_{K-1} - Q_K\|_{2,\mu} + \sqrt{C_{\mathrm{tv}}}\|\mathcal{T}Q_{K-1} - \mathcal{T}_{g_{K-1}}Q_{K-1}\|_{1,\mu},
\end{aligned}$$

where (d) follows by Assumption 9, (e) from Equation (30) and the fact $|\sup_x p(x) - \sup_x q(x)| \leq \sup_x |p(x) - q(x)|$, (f) from the fact $|(x)_+ - (y)_+| \leq |(x-y)_+|$, (g) follows by Jensen's inequality and by the facts $|\sup_x p(x) - \sup_x q(x)| \leq \sup_x |p(x) - q(x)|$ and $(x)_+ \leq |x|$, (h) follows by defining the distribution ν' as $\nu'(s',a') = \sum_{s,a}\nu(s,a)P^o_{s,a}(s')\mathbb{1}\{a' = \arg\max_b |Q^{\pi^*}(s',b) - Q_{K-1}(s',b)|\}$, and (i) using the fact that $\|\cdot\|_{1,\mu} \leq \|\cdot\|_{2,\mu}$. The rest of the proof follows similarly as in the proof of Theorem 1. □

F. Hybrid Robust φ-regularized RL Results

In this section, we set $V_{\max} = H$ whenever we use results from Proposition 3. We remark that we have attempted to optimize the absolute constants inside log factors of the performance guarantees. In the following, we use constants c_1, c_2, c_3 from Proposition 3.

Now we provide an extension of Proposition 7 using Proposition 4 when the data comes from adaptive sampling.

Proposition 9 (Online Dual Optimization Error Bound)**.** *Fix $\delta \in (0,1)$. For $k \in \{0,1,\cdots,K-1\}$, $h \in \{0,1,\cdots,H-1\}$, let g_h^k be the dual optimization function from Algorithm 2 (Step 4) for the state-action value function Q_{h+1}^k using samples in the dataset $\{\mathcal{D}_h^\mu, \mathcal{D}_h^0, \cdots, \mathcal{D}_h^{k-1}\}$. Let $\mathcal{T}_g$ be as defined in (14) and let $N = m_{\text{off}} + K \cdot m_{\text{on}}$. Then, with probability at least $1-\delta$, we have*

$$\|\mathcal{T}Q_{h+1}^k - \mathcal{T}_{g_h^k} Q_{h+1}^k\|_{1,\mu_h} \leq \frac{1}{m_{\text{off}}}\left(3\varepsilon_{\mathcal{G}} N + 48c_1 \log(2HK|\mathcal{G}||\mathcal{F}|/\delta)\right) = \Delta_{\text{dual,off}} \quad \textit{and}$$

$$\sum_{\tau=0}^{k-1} \|\mathcal{T}Q_{h+1}^k - \mathcal{T}_{g_h^k} Q_{h+1}^k\|_{1,d_h^{\pi_\tau}} \leq \frac{1}{m_{\text{on}}}\left(3\varepsilon_{\mathcal{G}} N + 48c_1 \log(2HK|\mathcal{G}||\mathcal{F}|/\delta)\right) = \Delta_{\text{dual,on}}.$$

Proof. Fix $k \in \{0,1,\cdots,K-1\}$, $h \in \{0,1,\cdots,H-1\}$, $Q_{h+1}^k \in \mathcal{F}_{h+1}$. The algorithm solves for g_h^k in the empirical risk minimization step as:

$$g_h^k = \arg\min_{g \in \mathcal{G}_h} \widehat{L}_{\text{dual}}(g; Q_{h+1}^k, \mathcal{D}),$$

where dataset $\mathcal{D} = \{(s_h^i, a_h^i, s_{h+1}^i)\}_{i \leq N}$ with $N = m_{\text{off}} + k \cdot m_{\text{on}}$. The first m_{off} samples in $\mathcal{D}$ are $\{(s_h^i, a_h^i, s_{h+1}^i)\}_{i \leq m_{\text{off}}} = \mathcal{D}_h^\mu$ (recall that these are generated by the offline state-action distribution μ_h), the next m_{on} samples are $\{(s_h^i, a_h^i, s_{h+1}^i)\}_{i=m_{\text{off}}+1}^{m_{\text{off}}+m_{\text{on}}} = \mathcal{D}_h^0$ (recall that these are generated by the state-action distribution $d_h^{\pi_0}$), and so on where the samples $\{(s_h^i, a_h^i, s_{h+1}^i)\}_{i=m_{\text{off}}+\tau \cdot m_{\text{on}}+1}^{m_{\text{off}}+(\tau+1)m_{\text{on}}} = \mathcal{D}_h^\tau$ (recall that these are generated by the state-action distribution $d_h^{\pi_\tau}$) for all $\tau \leq k-1$. We first have the following from step (b) in the proof of Proposition 7:

$$\begin{aligned}
&m_{\text{off}}\|\mathcal{T}Q_{h+1}^k - \mathcal{T}_{g_h^k} Q_{h+1}^k\|_{1,\mu} + m_{\text{on}} \sum_{\tau=0}^{k-1} \|\mathcal{T}Q_{h+1}^k - \mathcal{T}_{g_h^k} Q_{h+1}^k\|_{1,d_h^{\pi_\tau}} \\
&= m_{\text{off}}[\mathbb{E}_{s,a\sim\mu_h, s'\sim P_{s,a}^o}(\lambda\varphi^*((g_h^k(s,a) - \max_{a'} Q_{h+1}^k(s',a'))/\lambda) - g_h^k(s,a)) \\
&\qquad - \inf_{g \in L^1(\mu_h)} \mathbb{E}_{s,a\sim\mu_h, s'\sim P_{s,a}^o}(\lambda\varphi^*((g(s,a) - \max_{a'} Q_{h+1}^k(s',a'))/\lambda) - g(s,a))] \\
&\qquad + m_{\text{on}} \sum_{\tau=0}^{k-1} [\mathbb{E}_{s,a\sim d_h^{\pi_\tau}, s'\sim P_{s,a}^o}(\lambda\varphi^*((g_h^k(s,a) - \max_{a'} Q_{h+1}^k(s',a'))/\lambda) - g_h^k(s,a)) \\
&\qquad - \inf_{g \in L^1(d_h^{\pi_\tau})} \mathbb{E}_{s,a\sim d_h^{\pi_\tau}, s'\sim P_{s,a}^o}(\lambda\varphi^*((g(s,a) - \max_{a'} Q_{h+1}^k(s',a'))/\lambda) - g(s,a))] \\
&\overset{(a)}{=} m_{\text{off}}[\mathbb{E}_{s,a\sim\mu_h, s'\sim P_{s,a}^o}(\lambda\varphi^*((g_h^k(s,a) - \max_{a'} Q_{h+1}^k(s',a'))/\lambda) - g_h^k(s,a)) \\
&\qquad - \mathbb{E}_{s,a\sim\mu_h, s'\sim P_{s,a}^o}(\lambda\varphi^*((g_{-1}^*(s,a) - \max_{a'} Q_{h+1}^k(s',a'))/\lambda) - g_{-1}^*(s,a))] \\
&\qquad + m_{\text{on}} \sum_{\tau=0}^{k-1} [\mathbb{E}_{s,a\sim d_h^{\pi_\tau}, s'\sim P_{s,a}^o}(\lambda\varphi^*((g_h^k(s,a) - \max_{a'} Q_{h+1}^k(s',a'))/\lambda) - g_h^k(s,a)) \\
&\qquad - \mathbb{E}_{s,a\sim d_h^{\pi_\tau}, s'\sim P_{s,a}^o}(\lambda\varphi^*((g_\tau^*(s,a) - \max_{a'} Q_{h+1}^k(s',a'))/\lambda) - g_\tau^*(s,a))] \\
&= \sum_{i=1}^{m_{\text{off}}} \mathbb{E}_{s_h^i, a_h^i\sim\mu_h, s_{h+1}^i\sim P_{s_h^i,a_h^i}^o}[(\lambda\varphi^*((g_h^k(s_h^i,a_h^i) - \max_{a'} Q_{h+1}^k(s_{h+1}^i,a'))/\lambda) - g_h^k(s_h^i,a_h^i)) \\
&\qquad - (\lambda\varphi^*((g_{-1}^*(s_h^i,a_h^i) - \max_{a'} Q_{h+1}^k(s_{h+1}^i,a'))/\lambda) - g_{-1}^*(s_h^i,a_h^i))] \\
&\qquad + \sum_{i=m_{\text{off}}+1}^{m_{\text{off}}+m_{\text{on}}} \mathbb{E}_{s_h^i, a_h^i\sim d_h^{\pi_0}, s_{h+1}^i\sim P_{s_h^i,a_h^i}^o}[(\lambda\varphi^*((g_h^k(s_h^i,a_h^i) - \max_{a'} Q_{h+1}^k(s_{h+1}^i,a'))/\lambda) - g_h^k(s_h^i,a_h^i))
\end{aligned}$$

$$
\begin{aligned}
&- (\lambda\varphi^*((g_0^*(s_h^i, a_h^i) - \max_{a'} Q_{h+1}^k(s_{h+1}^i, a'))/\lambda) - g_0^*(s_h^i, a_h^i))] \\
&+ \cdots \\
\overset{(b)}{\leq}\ & 3\varepsilon_{\mathcal{G}} N + 48 c_1 \log(2|\mathcal{G}||\mathcal{F}|/\delta),
\end{aligned}
$$

where (a) follows by defining the corresponding true solutions g_τ^* for all $\tau \in \{-1, 0, 1, \cdots, k-1\}$. For (b) with the empirical risk minimization solution g_h^k, we use Proposition 4 by setting $c = c_1$ (with c_1, constant dependent on H and λ, from Proposition 3) and since $g_h^k \in \mathcal{G}_h, Q_{h+1}^k \in \mathcal{F}_{h+1}$ with sizes $|\mathcal{G}_h| \leq |\mathcal{G}|$ and $|\mathcal{F}_{h+1}| \leq |\mathcal{F}|$ under the union bound. Taking a union bound over $k \in \{0, 1, \cdots, K-1\}$, $h \in \{0, 1, \cdots, H-1\}$, and bounding each term separately, completes the proof. □

Now we provide an extension of Proposition 8 using Lemma 7 when the data comes from adaptive sampling.

Proposition 10 (Online Least-squares Generalization Bound)**.** *Fix $\delta \in (0, 1)$. For $k \in \{0, 1, \cdots, K-1\}$, $h \in \{0, 1, \cdots, H-1\}$, let Q_h^k be the least-squares solution from Algorithm 2 (Step 5) for the state-action value function Q_{h+1}^k and dual variable function g_h^k using samples in the dataset $\{\mathcal{D}_h^\mu, \mathcal{D}_h^0, \cdots, \mathcal{D}_h^{k-1}\}$. Let $\mathcal{T}_g$ be as defined in (14) and let $N = m_{\text{off}} + K \cdot m_{\text{on}}$. Then, with probability at least $1-\delta$, we have*

$$
\|\mathcal{T}_{g_h^k} Q_{h+1}^k - Q_h^k\|_{2,\mu_h} \leq \frac{1}{\sqrt{m_{\text{off}}}} \left(\sqrt{3\varepsilon_{\mathcal{F},\text{r}} N} + 8(1 + c_1 + H)\sqrt{\log(2HK|\mathcal{G}||\mathcal{F}|/\delta)} \right) = \Delta_{\text{rQ,off}} \quad \textit{and}
$$

$$
\sqrt{\sum_{\tau=0}^{k-1} \|\mathcal{T}_{g_h^k} Q_{h+1}^k - Q_h^k\|_{2,d_h^{\pi_\tau}}^2} \leq \frac{1}{\sqrt{m_{\text{on}}}} \left(\sqrt{3\varepsilon_{\mathcal{F},\text{r}} N} + 8(1 + c_1 + H)\sqrt{\log(2HK|\mathcal{G}||\mathcal{F}|/\delta)} \right) = \Delta_{\text{rQ,on}}.
$$

Proof. We adapt the proof of Song et al. (2023, Lemma 7) here. Fix $k \in \{0, 1, \cdots, K-1\}$, $h \in \{0, 1, \cdots, H-1\}$, $g_h^k \in \mathcal{G}_h$, and $Q_{h+1}^k \in \mathcal{F}_{h+1}$. The algorithm solves for Q_h^k in the least-squares regression step as:

$$
Q_h^k = \arg\min_{Q \in \mathcal{F}_h} \widehat{L}_{\text{robQ}}(Q; Q_{h+1}^k, g_h^k, \mathcal{D}),
$$

where dataset $\mathcal{D} = \{(x_i, y_i)\}_{i \leq N}$ with $N = m_{\text{off}} + k \cdot m_{\text{on}}$ and

$$
x_i = (s_h^i, a_h^i) \qquad \text{and} \qquad y_i = r_h(s_h^i, a_h^i) - \lambda\varphi^*((g_h^k(s_h^i, a_h^i) - \max_{a'} Q_{h+1}^k(s_{h+1}^i, a'))/\lambda) + g_h^k(s_h^i, a_h^i).
$$

The first m_{off} samples in $\mathcal{D}$ are $\{(x_i, y_i)\}_{i \leq m_{\text{off}}} = \mathcal{D}_h^\mu$ (recall that these are generated by the offline state-action distribution μ_h), the next m_{on} samples are $\{(x_i, y_i)\}_{i=m_{\text{off}}+1}^{m_{\text{off}}+m_{\text{on}}} = \mathcal{D}_h^0$ (recall that these are generated by the state-action distribution $d_h^{\pi_0}$), and so on where the samples $\{(x_i, y_i)\}_{i=m_{\text{off}}+\tau \cdot m_{\text{on}}+1}^{m_{\text{off}}+(\tau+1)m_{\text{on}}} = \mathcal{D}_h^\tau$ (recall that these are generated by the state-action distribution $d_h^{\pi_\tau}$) for all $\tau \leq k-1$.

For using Lemma 7, we first note for any sample (x, y) in $\mathcal{D}$ with $x = (s_h, a_h)$ and $y = (r_h(s_h, a_h) - \lambda\varphi^*((g_h^k(s_h, a_h) - \max_{a' \in \mathcal{A}_{h+1}} Q_{h+1}^k(s_{h+1}, a'))/\lambda) + g_h^k(s_h, a_h))$, there exists some $f_{h+1} \in \mathcal{F}_{h+1}$ by Assumption 5 such that the following holds:

$$
\begin{aligned}
\mathbb{E}[y \mid x] &= \mathbb{E}_{s_{h+1} \sim P_{h,s_h,a_h}^o} (r_h(s_h, a_h) - \lambda\varphi^*((g_h^k(s_h, a_h) - \max_{a' \in \mathcal{A}_{h+1}} Q_{h+1}^k(s_{h+1}, a'))/\lambda) + g_h^k(s_h, a_h)) \\
&= \mathcal{T}_{g_h^k} Q_{h+1}^k(s_h, a_h) \leq f_{h+1}(s_h, a_h).
\end{aligned}
$$

We also note for any sample in $\mathcal{D}$, $|y| \leq 1 + c_1$ (with c_1, constant dependent on H and λ, from Proposition 3) and $f_{h+1}(s, a) \leq H$ for all s, a. With these notes, applying Lemma 7, we get that the least square regression solution Q_h^k satisfies

$$
\sum_{i=1}^N \mathbb{E}[(\mathcal{T}_{g_h^k} Q_{h+1}^k(x_i) - Q_h^k(x_i))^2 \mid \mathcal{D}] \leq 3\varepsilon_{\mathcal{F},\text{r}} N + 64(1 + c_1 + H)^2 \log(2|\mathcal{G}||\mathcal{F}|/\delta)
$$

with probability at least $1-\delta$, since $g_h^k \in \mathcal{G}_h$ and $Q_{h+1}^k \in \mathcal{F}_{h+1}$ with sizes $|\mathcal{G}_h| \le |\mathcal{G}|$ and $|\mathcal{F}_{h+1}| \le |\mathcal{F}|$ under the union bound. Recall the samples in $\mathcal{D}_h^\mu$ are independently and identically drawn from the offline distribution μ_h, and the samples in $\mathcal{D}_h^\tau$ are independently and identically drawn from the state-action distribution $d_h^{\pi_\tau}$. Thus we can further write as

$$m_{\text{off}}\|\mathcal{T}_{g_h^k}Q_{h+1}^k - Q_h^k\|_{2,\mu}^2 + m_{\text{on}}\sum_{\tau=0}^{k-1}\|\mathcal{T}_{g_h^k}Q_{h+1}^k - Q_h^k\|_{2,d_h^{\pi_\tau}}^2 \le 3\varepsilon_{\mathcal{F},\text{r}}N + 64(1+c_1+H)^2\log(2|\mathcal{G}||\mathcal{F}|/\delta).$$

Taking a union bound over $k \in \{0,1,\cdots,K-1\}$, $h \in \{0,1,\cdots,H-1\}$, bounding each term separately, and using the fact $\sqrt{x+y} \le \sqrt{x}+\sqrt{y}$, completes the proof. □

We are now ready to prove the main theorem.

F.1. Proof of Theorem 2

Theorem 5 (Restatement of Theorem 2)**.** *Let Assumptions 4 to 8 hold and fix any $\delta \in (0,1)$. Then, HyTQ algorithm policies $\{\pi_k\}_{k\in[K]}$ satisfy*

$$\sum_{k=0}^{K-1}(V^{\pi^*} - V^{\pi_k}) \le \mathcal{O}((\sqrt{\varepsilon_{\mathcal{F},\text{r}}} + \varepsilon_{\mathcal{G}})K^{5/2}H) + \widetilde{\mathcal{O}}(\max\{C(\pi^*),1\}\sqrt{dKH^2}(\lambda+H)\log(HK|\mathcal{F}||\mathcal{G}|/\delta)\sqrt{\log(1+(K/d))})$$

with probability at least $1-\delta$.

Proof. We let $V_h^k(s) = Q_h^k(s,\pi_k(s))$ for every s,h. Since π_k is the greedy policy w.r.t Q^k, we also have $V_h^k(s) = Q_h^k(s,\pi_k(s)) = \max_a Q_h^k(s,a)$. We recall that $V^* = V^{\pi^*}$ and $Q^* = Q^{\pi^*}$. We also note that the same holds true for any stationary Markov policy π from (Zhang et al., 2024) that Q^π satisfies $Q_h^\pi(s,a) = r_h(s,a) + \gamma \min_{P_{h,s,a} \ll P_{h,s,a}^o}(\mathbb{E}_{s'\sim P_{h,s,a}}[V_h^\pi(s')] + \lambda D_\varphi(P_{h,s,a}, P_{h,s,a}^o))$. We can now further use the dual form (4) under Assumption 8, that is, for all π and $f_{h+1} \in \mathcal{F}_{h+1}$,

$$\begin{aligned}
Q_h^\pi(s,a) &= r_h(s,a) - \inf_{\eta\in[0,\lambda]}\left(\mathbb{E}_{s'\sim P_{h,s,a}^o}[(\eta - V_{h+1}^\pi(s'))_+] - \eta\right), \text{ and} \\
(\mathcal{T}f_{h+1})(s,a) &= r_h(s,a) - \inf_{\eta\in[0,\lambda]}\left(\mathbb{E}_{s'\sim P_{h,s,a}^o}[(\eta - \max_{a'} f_{h+1}(s',a'))_+] - \eta\right) \\
(\mathcal{T}_{g_h}f_{h+1})(s,a) &= r_h(s,a) - \mathbb{E}_{s'\sim P_{h,s,a}^o}[(g_h(s,a) - \max_{a'} f_{h+1}(s',a'))_+] + g_h(s,a).
\end{aligned} \tag{33}$$

We first characterize the performance decomposition between $V_0^{\pi^*}$ and $V_0^{\pi_k}$. We recall the initial state distribution d_0. Since $V^{\pi^*}(s) \ge V^{\pi_k}(s)$ for any $s \in \mathcal{S}$, we observe that

$$\begin{aligned}
0 &\le \sum_{k=0}^{K-1}\mathbb{E}_{s_0\sim d_0}[V_0^{\pi^*}(s_0) - V_0^{\pi_k}(s_0)] = \sum_{k=0}^{K-1}\mathbb{E}_{s_0\sim d_0}[(V_0^{\pi^*}(s_0) - V_0^k(s_0)) - (V_0^{\pi_k}(s_0) - V_0^k(s_0))] \\
&= \sum_{k=0}^{K-1}\mathbb{E}_{s_0\sim d_0}[(Q_0^{\pi^*}(s_0,\pi^*(s_0)) - Q_0^k(s_0,\pi_k(s_0))) - (Q_0^{\pi_k}(s_0,\pi_k(s_0)) - Q_0^k(s_0,\pi_k(s_0)))] \\
&\le \underbrace{\sum_{k=0}^{K-1}\mathbb{E}_{s_0\sim d_0}[(Q_0^{\pi^*}(s_0,\pi^*(s_0)) - Q_0^k(s_0,\pi_k(s_0)))_+]}_{(I)} + \underbrace{\sum_{k=0}^{K-1}\mathbb{E}_{s_0\sim d_0}[(Q_0^k(s_0,\pi_k(s_0)) - Q_0^{\pi_k}(s_0,\pi_k(s_0)))_+]}_{(II)}.
\end{aligned} \tag{34}$$

We rewrite the state-action distribution $d_{P^o}^{h,\pi}$, dropping P^o, as d_h^π for simplicity. Letting d_h^π also denote a state distribution ($\Delta(\mathcal{S})$), we can write it as, for all h,

$$d_h^\pi = \begin{cases} d_0 & \text{if } h = 0, \\ P_{h,s',a'}^o & \text{otherwise, with } s' \sim d_{h-1}^\pi, a' \sim \pi_h(s'). \end{cases} \tag{35}$$

Analyzing one term in (I) of (34) starting with the facts that π_k is the greedy policy with respect to Q^k and function $(x)_+$ is non-decreasing in $x \in \mathbb{R}$:

$$\begin{aligned}
&\mathbb{E}_{s_0\sim d_0}[(Q_0^{\pi^*}(s_0,\pi^*(s_0)) - Q_0^k(s_0,\pi_k(s_0)))_+] \le \mathbb{E}_{s_0,a_0\sim d_0^{\pi^*}}[(Q_0^{\pi^*}(s_0,a_0) - Q_0^k(s_0,a_0))_+] \\
&\overset{(a)}{\le} \mathbb{E}_{s_0,a_0\sim d_0^{\pi^*}}[(Q_0^{\pi^*}(s_0,a_0) - \mathcal{T}Q_1^k(s_0,a_0))_+] + \mathbb{E}_{s_0,a_0\sim d_0^{\pi^*}}[(\mathcal{T}Q_1^k(s_0,a_0) - Q_0^k(s_0,a_0))_+] \\
&\overset{(b)}{\le} \mathbb{E}_{s_0,a_0\sim d_0^{\pi^*}}(\sup_{\eta}(\mathbb{E}_{s_1\sim P^o_{0,s_0,a_0}}[(\eta - \max_{a'} Q_1^k(s_1,a'))_+ - (\eta - \max_{a'} Q_1^{\pi^*}(s_1,a'))_+]))_+ \\
&\qquad\qquad + \mathbb{E}_{s_0,a_0\sim d_0^{\pi^*}}[(\mathcal{T}Q_1^k(s_0,a_0) - Q_0^k(s_0,a_0))_+] \\
&\overset{(c)}{\le} \mathbb{E}_{s_0,a_0\sim d_0^{\pi^*}}(\mathbb{E}_{s_1\sim P^o_{0,s_0,a_0}}(\max_{a'} Q_1^{\pi^*}(s_1,a') - \max_{a'} Q_1^k(s_1,a'))_+)_+ + \mathbb{E}_{s_0,a_0\sim d_0^{\pi^*}}[(\mathcal{T}Q_1^k(s_0,a_0) - Q_0^k(s_0,a_0))_+] \\
&\overset{(d)}{\le} \mathbb{E}_{s_0,a_0\sim d_0^{\pi^*}}\mathbb{E}_{s_1\sim P^o_{0,s_0,a_0}}(Q_1^{\pi^*}(s_1,\pi^*(s_1)) - Q_1^k(s_1,\pi_k(s_1)))_+ + \mathbb{E}_{s_0,a_0\sim d_0^{\pi^*}}[(\mathcal{T}Q_1^k(s_0,a_0) - Q_0^k(s_0,a_0))_+] \\
&= \mathbb{E}_{s_0\sim d_1^{\pi^*}}[(Q_1^{\pi^*}(s_1,\pi^*(s_1)) - Q_1^k(s_1,\pi_k(s_1)))_+] + \mathbb{E}_{s_0,a_0\sim d_0^{\pi^*}}[(\mathcal{T}Q_1^k(s_0,a_0) - Q_0^k(s_0,a_0))_+],
\end{aligned} \tag{36}$$

where (a) follows by triangle inequality for $(\cdot)_+$ operation, (b) from Bellman equation, operator $\mathcal{T}$, and the fact $\inf_x p(x) - \inf_x q(x) \le \sup_x(p(x) - q(x))$, (c) from the fact $(x)_+ - (y)_+ \le (x-y)_+$ for any $x, y \in \mathbb{R}$, (d) follows by Jensen's inequality and by definitions of policies π^* and π_k. Now, recursively applying this method for first term over horizon in (36) we get

$$\begin{aligned}
&\mathbb{E}_{s_0\sim d_0}[(Q_0^{\pi^*}(s_0,\pi^*(s_0)) - Q_0^k(s_0,\pi_k(s_0)))_+] \\
&\le \mathbb{E}_{s_H\sim d_H}[(Q_H^{\pi^*}(s_H,\pi^*(s_H)) - Q_H^k(s_H,\pi_k(s_H)))_+] + \sum_{h=0}^{H-1}\mathbb{E}_{s,a\sim d_h^{\pi^*}}[(\mathcal{T}Q_{h+1}^k(s,a) - Q_h^k(s,a))_+] \\
&\le \sum_{h=0}^{H-1}\mathbb{E}_{s,a\sim d_h^{\pi^*}}[(\mathcal{T}Q_{h+1}^k(s,a) - Q_h^k(s,a))_+],
\end{aligned} \tag{37}$$

where the last inequality holds since $V_H^\pi(s_H) = 0$ for all π and $Q_H^k(s_H,\pi_k(s_H)) = 0$.

Recall

$$C(\pi^*) = \max_{f\in\mathcal{F}} \frac{\sum_{h=0}^{H-1}\mathbb{E}_{s,a\sim d_h^{\pi^*}}[(\mathcal{T}f_{h+1}(s,a) - f_h(s,a))_+]}{\sum_{h=0}^{H-1}\mathbb{E}_{s,a\sim \mu_h}[|\mathcal{T}f_{h+1}(s,a) - f_h(s,a)|]}.$$

Now, using (37) in (I) of (34), the following holds with probability at least $1-\delta/2$:

$$\begin{aligned}
\sum_{k=0}^{K-1}\mathbb{E}_{s_0\sim d_0}[(Q^{\pi^*}(s_0,\pi^*(s_0)) - Q^k(s_0,\pi_k(s_0)))_+] &\le \sum_{k=0}^{K-1}\sum_{h=0}^{H-1}\mathbb{E}_{s,a\sim d_h^{\pi^*}}[(\mathcal{T}Q_{h+1}^k(s,a) - Q_h^k(s,a))_+] \\
&\overset{(e)}{\le} \sum_{k=0}^{K-1} C(\pi^*)\sum_{h=0}^{H-1}\|\mathcal{T}Q_{h+1}^k - Q_h^k\|_{1,\mu_h} \\
&\overset{(f)}{\le} \sum_{k=0}^{K-1} C(\pi^*)\sum_{h=0}^{H-1}(\|\mathcal{T}Q_{h+1}^k - \mathcal{T}_{g_h^k}Q_{h+1}^k\|_{1,\mu_h} + \|\mathcal{T}_{g_h^k}Q_{h+1}^k - Q_h^k\|_{2,\mu_h}) \\
&\overset{(g)}{\le} KHC(\pi^*)(\Delta_{\text{dual,off}} + \Delta_{\text{rQ,off}}),
\end{aligned} \tag{38}$$

where (e) follows from definition of $C(\pi^*)$ in Assumption 4, (f) from triangle inequality and the fact $\|\cdot\|_{1,\mu} \le \|\cdot\|_{2,\mu}$, and (g) follows from Propositions 9 and 10.

For (II), firstly we note $\mathbb{E}_{s_0\sim d_0}[(Q^k(s_0,\pi_k(s_0)) - Q^{\pi_k}(s_0,\pi_k(s_0)))_+] = \mathbb{E}_{s_0,a_0\sim d_0^{\pi_k}}[(Q^k(s_0,a_0) - Q^{\pi_k}(s_0,a_0))_+]$. So, following the same analysis as in (I), we get

$$\mathbb{E}_{s_0\sim d_0}[(Q^k(s_0,\pi_k(s_0)) - Q^{\pi_k}(s_0,\pi_k(s_0)))_+] \le \sum_{h=0}^{H-1}\mathbb{E}_{s,a\sim d_h^{\pi_k}}[(Q_h^k(s,a) - \mathcal{T}Q_{h+1}^k(s,a))_+]$$

$$\leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi_k}}[(Q_h^k(s,a) - (\mathcal{T}_{g_h^k} Q_{h+1}^k)(s,a))_+ + ((\mathcal{T}_{g_h^k} Q_{h+1}^k)(s,a) - (\mathcal{T} Q_{h+1}^k)(s,a))_+], \tag{39}$$

where the last inequality follows by triangle inequality for $(\cdot)_+$ operation.

Now, using (39) in (II) of (34), we have

$$\sum_{k=0}^{K-1} \mathbb{E}_{s_0\sim d_0}[(Q^k(s_0,\pi_k(s_0)) - Q^{\pi_k}(s_0,\pi_k(s_0)))_+] \leq$$
$$\sum_{k=0}^{K-1}\sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi_k}}[(Q_h^k(s,a) - (\mathcal{T}_{g_h^k} Q_{h+1}^k)(s,a))_+] + \sum_{k=0}^{K-1}\sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi_k}}[((\mathcal{T}_{g_h^k} Q_{h+1}^k)(s,a) - \mathcal{T} Q_{h+1}^k(s,a))_+]. \tag{40}$$

Recall bilinear model from Assumption 7: $\mathbb{E}_{d_h^{\pi^f}}[(f_h - \mathcal{T}_{g_h} f_{h+1})_+] = |\langle X_h(f), W_h^{\mathrm{q}}(f,g)\rangle|$.

Analyzing the first part of (40), the following holds with probability at least $1-\delta/2$:

$$\begin{aligned}
\sum_{k=0}^{K-1}\sum_{h=0}^{H-1} \mathbb{E}_{d_h^{\pi_k}}[(Q_h^k - \mathcal{T}_{g_h^k} Q_{h+1}^k)_+] &\overset{(h)}{=} \sum_{k=0}^{K-1}\sum_{h=0}^{H-1} \left|\left\langle X_h(Q^k), W_h^{\mathrm{q}}(Q^k,g^k)\right\rangle\right| \\
&\overset{(i)}{\leq} \sum_{k=0}^{K-1}\sum_{h=0}^{H-1} \|X_h(Q^k)\|_{\Sigma_{k-1;h}^{-1}} \|W_h^{\mathrm{q}}(Q^k,g^k)\|_{\Sigma_{k-1;h}} \\
&= \sum_{k=0}^{K-1}\sum_{h=0}^{H-1} \|X_h(Q^k)\|_{\Sigma_{k-1;h}^{-1}} \sqrt{(W_h^{\mathrm{q}}(Q^k,g^k))^\top \Sigma_{k-1;h} W_h^{\mathrm{q}}(Q^k,g^k)} \\
&= \sum_{k=0}^{K-1}\sum_{h=0}^{H-1} \|X_h(Q^k)\|_{\Sigma_{k-1;h}^{-1}} \sqrt{(W_h^{\mathrm{q}}(Q^k,g^k))^\top (\sum_{i=0}^{k-1} X_h(Q^i)X_h(Q^i)^\top + \sigma\mathbb{1}) W_h^{\mathrm{q}}(Q^k,g^k)} \\
&= \sum_{k=0}^{K-1}\sum_{h=0}^{H-1} \|X_h(Q^k)\|_{\Sigma_{k-1;h}^{-1}} \sqrt{\sum_{i=0}^{k-1} |\langle W_h^{\mathrm{q}}(Q^k,g^k), X_h(Q^i)\rangle|^2 + \sigma \|W_h^{\mathrm{q}}(Q^k,g^k)\|^2} \\
&\overset{(j)}{\leq} \sum_{k=0}^{K-1}\sum_{h=0}^{H-1} \|X_h(Q^k)\|_{\Sigma_{k-1;h}^{-1}} \sqrt{\sum_{i=0}^{k-1} |\langle W_h^{\mathrm{q}}(Q^k,g^k), X_h(Q^i)\rangle|^2 + \sigma B_W^2} \\
&\overset{(k)}{\leq} \sum_{k=0}^{K-1}\sum_{h=0}^{H-1} \|X_h(Q^k)\|_{\Sigma_{k-1;h}^{-1}} \sqrt{\sum_{i=0}^{k-1} \|\mathcal{T}_{g_h^k} Q_{h+1}^k - Q_h^k\|_{2,d_h^{\pi_i}}^2 + \sigma B_W^2} \\
&\overset{(l)}{\leq} \sum_{k=0}^{K-1}\sum_{h=0}^{H-1} \|X_h(Q^k)\|_{\Sigma_{k-1;h}^{-1}} (\sqrt{\sum_{i=0}^{k-1} \|\mathcal{T}_{g_h^k} Q_{h+1}^k - Q_h^k\|_{2,d_h^{\pi_i}}^2} + \sqrt{\sigma B_W^2}) \\
&\overset{(m)}{\leq} (\Delta_{\mathrm{rQ,on}} + \sqrt{\sigma B_W^2}) \sum_{k=0}^{K-1}\sum_{h=0}^{H-1} \|X_h(Q^k)\|_{\Sigma_{k-1;h}^{-1}} \\
&\overset{(n)}{\leq} (\Delta_{\mathrm{rQ,on}} + B_X B_W)\sqrt{2dH^2 \log(1+\frac{K}{d})K},
\end{aligned} \tag{41}$$

where (h) follows from Assumption 7, (i) from matrix Cauchy-Schwarz inequality, (j) from Assumption 7, and (k) by Assumption 7 with $\|\cdot\|_{1,d_h^{\pi_i}} \leq \|\cdot\|_{2,d_h^{\pi_i}}$:

$$|\left\langle W_h^{\mathrm{q}}(Q^k,g^k), X_h(Q^i)\right\rangle| = \mathbb{E}_{s,a\sim d_h^{\pi_i}}[(Q_h^k(s,a) - (\mathcal{T} Q_{h+1}^k)(s,a))_+] \leq \|\mathcal{T}_{g_h^k} Q_{h+1}^k - Q_h^k\|_{2,d_h^{\pi_i}}.$$

Finally, (l) follows by the fact $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$, (m) follows from Proposition 10, and (n) follows from Lemma 6.

Now recall bilinear model from Assumption 7: $\mathbb{E}_{d_h^{\pi^f}}[(\mathcal{T}_{g_h} f_{h+1} - \mathcal{T} f_{h+1})_+] = \left|\left\langle X_h(f), W_h^{\mathrm{d}}(f,g)\right\rangle\right|$. Following analysis above in (41) for the second part of (40) using Assumption 7 and Proposition 9, the following holds with probability at least $1-\delta/2$:

$$\sum_{k=0}^{K-1}\sum_{h=0}^{H-1}\mathbb{E}_{s,a\sim d_h^{\pi_k}}[(\mathcal{T}_{g_h^k} Q_{h+1}^k - \mathcal{T} Q_{h+1}^k)_+] \le (\Delta_{\mathrm{dual,on}} + B_X B_W)\sqrt{2dH^2\log(1+\frac{K}{d})K}. \tag{42}$$

Now combining Equations (41) and (42) with (40) we have

$$\sum_{k=0}^{K-1}\sum_{h=0}^{H-1}\mathbb{E}_{s,a\sim d_h^{\pi_k}}[(Q^k(s_0,\pi_k(s_0)) - Q^{\pi_k}(s_0,\pi_k(s_0)))_+]$$
$$\le (\Delta_{\mathrm{dual,on}} + \Delta_{\mathrm{rQ,on}} + 2B_X B_W)\sqrt{2dH^2\log(1+\frac{K}{d})K},$$

with probability at least $1-\delta$. Finally, we combine this and (38) with (34):

$$0 \le \sum_{k=0}^{K-1}\mathbb{E}_{s_0\sim d_0}[V_0^{\pi^*}(s_0) - V_0^{\pi_k}(s_0)] \le KHC(\pi^*)(\Delta_{\mathrm{dual,off}} + \Delta_{\mathrm{rQ,off}}) +$$
$$(\Delta_{\mathrm{dual,on}} + \Delta_{\mathrm{rQ,on}} + 2B_X B_W)\sqrt{2dH^2\log(1+\frac{K}{d})K}.$$

Let $N = m_{\mathrm{off}} + K\cdot m_{\mathrm{on}}$. Using offline bounds from Propositions 9 and 10 with $c_1 = 2\lambda + H$ from Proposition 3, we have:

$$0 \le \sum_{k=0}^{K-1}\mathbb{E}_{s_0\sim d_0}[V_0^{\pi^*}(s_0) - V_0^{\pi_k}(s_0)] \le KHC(\pi^*)\cdot$$
$$(\frac{1}{m_{\mathrm{off}}}(3\varepsilon_{\mathcal{G}}N + 48(2\lambda+H)\log(2HK|\mathcal{G}||\mathcal{F}|/\delta)) + \frac{1}{\sqrt{m_{\mathrm{off}}}}\left(\sqrt{3\varepsilon_{\mathcal{F},\mathrm{r}}N} + 8(1+2\lambda+2H)\sqrt{\log(2HK|\mathcal{G}||\mathcal{F}|/\delta)}\right))$$
$$+ (\Delta_{\mathrm{dual,on}} + \Delta_{\mathrm{rQ,on}} + 2B_X B_W)\sqrt{2dH^2\log(1+\frac{K}{d})K}.$$

Now using on-policy bounds from Propositions 9 and 10 with $c_1 = 2\lambda + H$ from Proposition 3, we have:

$$0 \le \sum_{k=0}^{K-1}\mathbb{E}_{s_0\sim d_0}[V_0^{\pi^*}(s_0) - V_0^{\pi_k}(s_0)] \le KHC(\pi^*)\cdot$$
$$(\frac{1}{m_{\mathrm{off}}}(3\varepsilon_{\mathcal{G}}N + 48(2\lambda+H)\log(2HK|\mathcal{G}||\mathcal{F}|/\delta)) + \frac{1}{\sqrt{m_{\mathrm{off}}}}\left(\sqrt{3\varepsilon_{\mathcal{F},\mathrm{r}}N} + 8(1+2\lambda+2H)\sqrt{\log(2HK|\mathcal{G}||\mathcal{F}|/\delta)}\right))$$
$$+ (\frac{1}{m_{\mathrm{on}}}(3\varepsilon_{\mathcal{G}}N + 48(2\lambda+H)\log(2HK|\mathcal{G}||\mathcal{F}|/\delta))$$
$$+ \frac{1}{\sqrt{m_{\mathrm{on}}}}\left(\sqrt{3\varepsilon_{\mathcal{F},\mathrm{r}}N} + 8(1+2\lambda+2H)\sqrt{\log(2HK|\mathcal{G}||\mathcal{F}|/\delta)}\right) + 2B_X B_W)\cdot\sqrt{2dH^2\log(1+\frac{K}{d})K}$$

Finally, choosing higher order terms by setting $m_{\mathrm{on}} = 1$ and $m_{\mathrm{off}} = K$, we get

$$0 \le \sum_{k=0}^{K-1}\mathbb{E}_{s_0\sim d_0}[V_0^{\pi^*}(s_0) - V_0^{\pi_k}(s_0)]$$
$$\le \sqrt{K}HC(\pi^*)(6(\varepsilon_{\mathcal{G}} + \sqrt{\varepsilon_{\mathcal{F},\mathrm{r}}})K^2 + (8 + 112\lambda + 64H)\log(2HK|\mathcal{G}||\mathcal{F}|/\delta))$$
$$+ (6(\varepsilon_{\mathcal{G}} + \sqrt{\varepsilon_{\mathcal{F},\mathrm{r}}})K^2 + 8 + 112\lambda + 64H\log(2HK|\mathcal{G}||\mathcal{F}|/\delta) + 2B_X B_W)\cdot\sqrt{2dH^2\log(1+\frac{K}{d})K}$$
$$\le \mathcal{O}((\sqrt{\varepsilon_{\mathcal{F},\mathrm{r}}} + \varepsilon_{\mathcal{G}})K^{5/2}H) + \widetilde{\mathcal{O}}(\max\{C(\pi^*),1\}\sqrt{dKH^2}(\lambda+H)\log(HK|\mathcal{F}||\mathcal{G}|/\delta)\sqrt{\log(1+(K/d))}).$$

The proof is now complete. □

F.2. HyTQ Algorithm Specialized Results

In this section we specialize our main result Theorem 2 for different bilinear model classes and also provide an equivalent sample complexity guarantee in the offline robust RL setting.

Before we move ahead, we showcase an important property of our robust transfer coefficient $C(\pi)$ for any fixed policy. Fixing a nominal model P^o, the transfer coefficient considers the distribution shift w.r.t the data-generating distribution along the general function class which the algorithm uses. It is in fact smaller than the existing density ratio based concentrability assumption (Assumption 9). We state this result in the following lemma.

Lemma 8. *For any policy π and offline distribution μ, we have $C(\pi) \leq \sup_{h,s,a} d_h^\pi(s,a)/\mu_h(s,a)$.*

Proof. By definition in Assumption 4, we get that

$$\begin{aligned}
C(\pi) &= \max_{f\in\mathcal{F}} \frac{\sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^\pi}[(\mathcal{T}f_{h+1}(s,a) - f_h(s,a))_+]}{\sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim \mu_h}[|\mathcal{T}f_{h+1}(s,a) - f_h(s,a)|]} \\
&\leq \max_{f\in\mathcal{F}} \frac{\sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^\pi}[|\mathcal{T}f_{h+1}(s,a) - f_h(s,a)|]}{\sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim \mu_h}[|\mathcal{T}f_{h+1}(s,a) - f_h(s,a)|]} \\
&\overset{(a)}{\leq} \max_{f\in\mathcal{F},h\in[H]} \frac{\mathbb{E}_{s,a\sim d_h^\pi}[|\mathcal{T}f_{h+1}(s,a) - f_h(s,a)|]}{\mathbb{E}_{s,a\sim \mu_h}[|\mathcal{T}f_{h+1}(s,a) - f_h(s,a)|]} \leq \sup_{h,s,a} \frac{d_h^\pi(s,a)}{\mu_h(s,a)},
\end{aligned}$$

where (a) follows from the Mediant inequality. □

Remark 5. *The concentrability assumption (Assumption 9) is in fact the same non-robust RL concentrability assumption (Munos & Szepesvári, 2008; Chen & Jiang, 2019). We make two important points here. Firstly, our transfer coefficient is larger than the transfer coefficient (Song et al., 2023, Definition 1) using the fact $\|\cdot\|_{1,\mu} \leq \|\cdot\|_{2,\mu}$. Secondly, our transfer coefficient is not directly comparable with the l2-norm version transfer coefficient (Xie et al., 2021, Definition 1). It is an interesting open question for future research to investigate about minimax lower bound guarantees w.r.t different transfer coefficients for both non-robust and robust RL problems.*

We now define a bilinear model called **Low Occupancy Complexity** (Du et al., 2021, Definition 4.7). The nominal model P^o and realizable function class $\mathcal{F}$ has *low occupancy complexity* w.r.t., for each $h \in [H]$, a (possibly unknown to the learner) feature map $\psi = (\psi_h : \mathcal{S}\times\mathcal{A} \to \mathcal{Y})$, where $\mathcal{Y}$ is a Hilbert space, and w.r.t. to a (possibly unknown to the learner) map $\nu_h : \mathcal{F} \mapsto \mathcal{Y}$ such that for all $f \in \mathcal{F}$, with greedy policy π^f w.r.t. f, and (s,a) we have

$$d_{P^o}^{h,\pi^f}(s,a) = \langle \psi_h(s,a), \nu_h(f)\rangle. \tag{43}$$

We make the following assumption on the offline data-generating distribution (or policy by slight notational override for convenience).

Assumption 11. *Consider the Low Occupancy Complexity model (bilinear model) on $\mathcal{Y} = \mathbb{R}^d$. Let the offline data distribution $\mu = \{\mu_h\}_{h\in[H]}$ satisfy a low rank structure, i.e. $\mu_h(s,a) = \langle \psi_h(s,a), \nu_h(f^{\text{off}})\rangle = \sum_{i\in[d]} \psi_{h,i}(s,a)\nu_{h,i}(f^{\text{off}})$, for some $f^{\text{off}} \in \mathcal{F}$.*

Now we extend our main result Theorem 2 in this next result specializing to the *Low Occupancy Complexity* (43) bilinear model.

Corollary 3 (Cumulative Suboptimality of Theorem 2 in Low Occupancy Complexity (43) bilinear model)**.** *Consider the Low Occupancy Complexity (43) bilinear model. Let Assumptions 4 to 6 and 8 hold and fix any $\delta \in (0,1)$. Then, HyTQ algorithm policies $\{\pi_k\}_{k\in[K]}$ satisfy*

$$\begin{aligned}
\sum_{k=0}^{K-1}(V^{\pi^*} - V^{\pi_k}) \leq& \mathcal{O}((\sqrt{\varepsilon_{\mathcal{F},\text{r}}} + \varepsilon_{\mathcal{G}})K^{5/2}H) \\
&+ \widetilde{\mathcal{O}}(\max\{C(\pi^*),1\}\sqrt{dKH^2}(\lambda + H)\log(HK|\mathcal{F}||\mathcal{G}|/\delta)\sqrt{\log(1+(K/d))}) \\
&+ \widetilde{\mathcal{O}}(\sqrt{dKH^4}\max_{f\in\mathcal{F}}\|\nu_h(f)\|_2\|\sum_{s,a}\psi_h(s,a)\|_2\sqrt{\log(1+(K/d))})
\end{aligned}$$

with probability at least $1-\delta$. Now, consider the offline data distribution as in Assumption 11 with perfect robust Bellman completeness, i.e. $\varepsilon_{\mathcal{F},\mathrm{r}} = 0 = \varepsilon_{\mathcal{G}}$. We have $C(\pi^*) \leq \sup_{h,i\in[d]}(\nu^*_{h,i}/\nu_{h,i}(f^{\mathrm{off}}))$.

Proof. Using the Low Occupancy Complexity (43) bilinear model, we have $\mathbb{E}_{d_{P^o}^{h,\pi^f}}[(\mathcal{T}_{g_h} f_{h+1} - \mathcal{T} f_{h+1})_+] = \left\langle X_h(f), W_h^{\mathrm{d}}(f,g)\right\rangle$, where

$$X_h(f) = \nu_h(f), \qquad W_h^{\mathrm{d}}(f,g) = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \psi_h(s,a)((\mathcal{T}_{g_h} f_{h+1})(s,a) - (\mathcal{T} f_{h+1})(s,a))_+.$$

We also have $\mathbb{E}_{d_{P^o}^{h,\pi^f}}[(f_h - \mathcal{T}_{g_h} f_{h+1})_+] = \langle X_h(f), W_h^{\mathrm{q}}(f,g)\rangle$, where

$$W_h^{\mathrm{q}}(f,g) = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \psi_h(s,a)(f_h(s,a) - (\mathcal{T}_{g_h} f_{h+1})(s,a))_+.$$

Furthermore, we set $B_X = \max_{f\in\mathcal{F}} \|\nu_h(f)\|_2$. Since $\mathcal{F}$ is realizable and $\mathcal{T}_g$ is complete, we set $B_W = H\|\sum_{s,a}\psi_h(s,a)\|_2$. Then the result directly follows by Theorem 2.

For the second statement, first note that the occupancy $d_h^{\pi^*}$ is low-rank as well since we assume perfect Bellman completeness. Following the proof of Lemma 8 we get

$$\begin{aligned}
C(\pi^*) &= \max_{f\in\mathcal{F}} \frac{\sum_{h=0}^{H-1}\mathbb{E}_{s,a\sim d_h^{\pi^*}}[(\mathcal{T} f_{h+1}(s,a) - f_h(s,a))_+]}{\sum_{h=0}^{H-1}\mathbb{E}_{s,a\sim\mu_h}[|\mathcal{T} f_{h+1}(s,a) - f_h(s,a)|]} \\
&\leq \max_{f\in\mathcal{F}} \frac{\sum_{h=0}^{H-1}\mathbb{E}_{s,a\sim d_h^{\pi^*}}[|\mathcal{T} f_{h+1}(s,a) - f_h(s,a)|]}{\sum_{h=0}^{H-1}\mathbb{E}_{s,a\sim\mu_h}[|\mathcal{T} f_{h+1}(s,a) - f_h(s,a)|]} \\
&\overset{(a)}{\leq} \max_{f\in\mathcal{F},h\in[H]} \frac{\mathbb{E}_{s,a\sim d_h^{\pi}}[|\mathcal{T} f_{h+1}(s,a) - f_h(s,a)|]}{\mathbb{E}_{s,a\sim\mu_h}[|\mathcal{T} f_{h+1}(s,a) - f_h(s,a)|]} \\
&\leq \sup_{h,s,a} \frac{d_h^{\pi}(s,a)}{\mu_h(s,a)} \overset{(b)}{\leq} \sup_{h,i\in[d]} \frac{\nu^*_{h,i}}{\nu_{h,i}(f^{\mathrm{off}})},
\end{aligned}$$

where $(a), (b)$ follows from the Mediant inequality. This completes the proof. □

We now define a bilinear model called **Low-rank Feature Selection Model** (Du et al., 2021, Definition A.1). The nominal model P^o is a *low-rank feature selection model* if it satisfies $P^o_{h,s,a}(s') = \langle\theta_h(s,a), \psi_h(s')\rangle$, for each $h\in[H]$ and all (s,a,s'), with a (possibly unknown to the learner) map $\theta = (\theta_h : \mathcal{S}\times\mathcal{A}\to\mathcal{Y})$ and a (possibly unknown to the learner) map $\psi_h : \mathcal{S}\mapsto\mathcal{Y}$, where $\mathcal{Y}$ is a Hilbert space.

This model specializes to the *kernel MDP model* when the map θ is known to the learner (Jin et al., 2021a, Definition 30). This model also specializes to the *low-rank MDP model* when $\mathcal{Y} = \mathbb{R}^d$ (Huang et al., 2023, Assumption 1) and furthermore to *linear MDP model* when the map θ is also known to the learner (Du et al., 2021, Definition A.4).

We make the following assumption on the offline data-generating distribution (or policy by slight notational override for convenience).

Assumption 12. *Consider the Low-rank MDP Model (bilinear model). Let the offline data distribution $\mu = \{\mu_h\}_{h\in[H]}$ satisfy $\max_{h,s,a}\pi_h^*(a|s)/\mu_h(a|s) \leq \alpha$ and suppose that μ is induced by the nominal model, i.e. $\mu_0(s) = d_0(s)$ (starting state distribution) and $\mu_h(s) = \mathbb{E}_{s',a'\sim\mu_{h-1}} P^o_{h-1,s',a'}(s)$ for any $h\geq 1$. Furthermore, suppose that μ satisfies that the feature covariance matrix $\Sigma_{\mu_{h-1},\theta} = \mathbb{E}_{s,a\sim\mu_{h-1}}[\theta_h(s,a)\theta_h(s,a)^\top]$ is invertible for all $h\in[H]$ and $\mathbb{E}_{s,a\sim\mu_h}[|\mathcal{T} f_{h+1}(s,a) - f_h(s,a)|] \geq 1$ for at least one $h\in[H]$ and all $f\in\mathcal{F}$.*

Now we extend our main result Theorem 2 in this next result specializing to the *Low-rank Feature Selection Model* bilinear model.

Corollary 4 (Cumulative Suboptimality of Theorem 2 in Low-rank Feature Selection Model (bilinear model))**.** *Consider the Low-rank Feature Selection Model (bilinear model). Let Assumptions 4 to 6 and 8 hold and fix any $\delta \in (0,1)$. Then, HyTQ algorithm policies $\{\pi_k\}_{k\in[K]}$ satisfy*

$$\begin{aligned}\sum_{k=0}^{K-1}(V^{\pi^*}-V^{\pi_k}) \leq& \mathcal{O}((\sqrt{\varepsilon_{\mathcal{F},\mathrm{r}}}+\varepsilon_{\mathcal{G}})K^{5/2}H)\\ &+\widetilde{\mathcal{O}}(\max\{C(\pi^*),1\}\sqrt{dKH^2}(\lambda+H)\log(HK|\mathcal{F}||\mathcal{G}|/\delta)\sqrt{\log(1+(K/d))})\\ &+\widetilde{\mathcal{O}}(\sqrt{dKH^4}\|\sum_{s,a}\theta_h(s,a)\|_2\|\sum_s\psi_h(s)\|_2\sqrt{\log(1+(K/d))})\end{aligned}$$

with probability at least $1-\delta$. Now, consider the offline data distribution as in Assumption 12 with a low-rank MDP model. We have

$$C(\pi^*) \leq \sqrt{2\alpha H}\sum_{h=1}^{H}\mathbb{E}_{s,a\sim d_{P^o}^{h-1,\pi^*}}\|\theta_h(s,a)\|_{\Sigma^{-1}_{\mu_{h-1},\theta}}+\sqrt{\alpha}.$$

Proof. We first begin with establishing a Q-value-dependent linearity property for the state-action-visitation measure $d_{P^o}^{h,\pi^f}(s,a)$. To do this, we adapt the proof of Huang et al. (2023, Lemma 17) here. We start by writing the state-visitation measure by recalling Equation (35) here:

$$\begin{aligned}d_{P^o}^{h,\pi^f}(s_h) &= \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} P^o_{h,s,a}(s_h)\pi^f_{h-1}(a|s)d_{P^o}^{h-1,\pi^f}(s)\\ &\overset{(a)}{=} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \langle\theta_h(s,a),\psi_h(s_h)\rangle\pi^f_{h-1}(a|s)d_{P^o}^{h-1,\pi^f}(s)\\ &= \langle\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\theta_h(s,a)\pi^f_{h-1}(a|s)d_{P^o}^{h-1,\pi^f}(s),\psi_h(s_h)\rangle = \langle\psi_h(s_h),\nu_{h,\pi^f}(f)\rangle,\end{aligned}$$

where (a) follows by the low-rank feature selection model definition, and the last equality follows by taking a functional $\nu_{h,\pi^f}(f)=\sum_{s,a}\theta_h(s,a)\pi^f_{h-1}(a|s)d_{P^o}^{h-1,\pi^f}(s)$. Since we consider the finite action space with possibly large state space setting for our results, the state-action visitation measure for the deterministic non-stationary policy π^f is now given by $d_{P^o}^{h,\pi^f}(s_h,a_h)=\langle\psi'_{h,\pi^f}(s_h,a_h),\nu_{h,\pi^f}(f)\rangle$ with $\psi'_{h,\pi^f}(s_h,a_h)=C\psi_h(s_h)1\{a_h=\pi^f_h(s)\}$ for features $\psi'_{h,\pi^f}:\mathcal{S}\times\mathcal{A}\to\mathcal{Y}$. Here $C>0$ is a normalizing constant such that the state-action visitation measure is a probability measure.

We now have $\mathbb{E}_{d_{P^o}^{h,\pi^f}}[(\mathcal{T}_{g_h}f_{h+1}-\mathcal{T}f_{h+1})_+]=\big\langle X_h(f),W_h^{\mathrm{d}}(f,g)\big\rangle$, where

$$X_h(f)=\nu_{h,\pi^f}(f),\qquad W_h^{\mathrm{d}}(f,g)=\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\psi'_{h,\pi^f}(s,a)((\mathcal{T}_{g_h}f_{h+1})(s,a)-(\mathcal{T}f_{h+1})(s,a))_+.$$

We also have $\mathbb{E}_{d_{P^o}^{h,\pi^f}}[(f_h-\mathcal{T}_{g_h}f_{h+1})_+]=\langle X_h(f),W_h^{\mathrm{q}}(f,g)\rangle$, where

$$W_h^{\mathrm{q}}(f,g)=\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\psi'_{h,\pi^f}(s,a)(f_h(s,a)-(\mathcal{T}_{g_h}f_{h+1})(s,a))_+.$$

Furthermore, we set

$$\max_{f\in\mathcal{F}}\|\nu_h(f)\|_2=\max_{f\in\mathcal{F}}\|\sum_{s,a}\theta_h(s,a)\pi^f(a|s)d_{P^o}^{h-1,\pi^f}(s)\|_2\leq\|\sum_{s,a}\theta_h(s,a)\|_2=B_X.$$

Since $\mathcal{F}$ is realizable and $\mathcal{T}_g$ is complete for all $g\in\mathcal{G}$, we set

$$H\|\sum_{s,a}\psi'_{h,\pi^f}(s,a)\|_2=HC\|\sum_{s,a}\psi_h(s)1\{a=\pi^f_h(s)\}\|_2\leq HC\|\sum_s\psi_h(s)\|_2=B_W.$$

Then the first result directly follows by Theorem 2. Following the proof of Song et al. (2023, Lemma 13) for our transfer coefficient $C(\pi^*)$, with the facts $(x-y)^2 \le |x-y||x+y|$ for $x, y \ge 0$ and $\|f_h\|_\infty \le H$ for all $h \in [H]$, the last statement for $C(\pi^*)$ follows. This completes the proof. □

Now we extend our main result Theorem 2 in this next result to showcase sample complexity for comparisons with offline+online RL setting.

Corollary 5 (Offline+Online RL Sample Complexity of the HyTQ algorithm)**.** *Let Assumptions 4 to 8 hold. Fix any $\delta \in (0,1)$ and any $\varepsilon > 0$, and let N_{tot} be the total number of sample tuples used in HyTQ algorithm. Then, the uniform policy $\widehat{\pi}$ (uniform convex combination) of HyTQ algorithm policies $\{\pi_k\}_{k\in[K]}$ satisfy, with probability at least $1-\delta$,*

$$V^{\pi^*} - V^{\widehat{\pi}} \le \varepsilon, \quad \text{if } N \ge N_{\text{tot}} = \widetilde{\mathcal{O}}\Big(\frac{\max\{(C(\pi^*))^2, 1\} d H^3 (\lambda + H)^2}{\varepsilon^2} \log^2(H|\mathcal{F}||\mathcal{G}|/\delta)\Big).$$

Proof. This proof is straightforward from the Theorem 2 using a standard online-to-batch conversion (Shalev-Shwartz & Ben-David, 2014, Theorem 14.8 & Chapter 21). Define the policy $\widehat{\pi} = \text{Uniform}\{\pi_0, \dots, \pi_{K-1}\}$. From Theorem 2, we get

$$\begin{aligned} 0 &\le \mathbb{E}_{s_0\sim d_0}[V_0^{\pi^*}(s_0) - V_0^{\widehat{\pi}}(s_0)] = \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}_{s_0\sim d_0}[V_0^{\pi^*}(s_0) - V_0^{\pi_k}(s_0)] \\ &\le \mathcal{O}((\sqrt{\varepsilon_{\mathcal{F},\text{r}}} + \varepsilon_{\mathcal{G}}) K^{3/2} H) + \widetilde{\mathcal{O}}(\max\{C(\pi^*), 1\}\sqrt{dH^2/K}(\lambda + H)\log(HK|\mathcal{F}||\mathcal{G}|/\delta)\sqrt{\log(1 + (K/d))}). \end{aligned}$$

We recall that our algorithm uses $m_{\text{off}} H$ number of offline samples and $m_{\text{on}} HK$ number of on-policy samples in the datasets $\{\mathcal{D}_h^\mu, \mathcal{D}_h^0, \cdots, \mathcal{D}_h^{K-1}\}$ for all $h \in [H]$. Since we set $m_{\text{on}} = 1$ and $m_{\text{off}} = K$, the total number of offline and on-policy samples is $2HK$.

Fix any $\varepsilon > 0$. For approximations $\varepsilon_{\mathcal{F},\text{r}}, \varepsilon_{\mathcal{G}}$, we first assume there exists $K_1 = \widetilde{\mathcal{O}}(H^4)$ such that $\mathcal{O}((\sqrt{\varepsilon_{\mathcal{F},\text{r}}} + \varepsilon_{\mathcal{G}}) K^{3/2} H) \le \varepsilon/2$ for all $K \ge K_1$. Let

$$K_2 = \widetilde{\mathcal{O}}\Big(\frac{\max\{(C(\pi^*))^2, 1\} d H^2 (\lambda + H)^2}{\varepsilon^2} \log^2(H|\mathcal{F}||\mathcal{G}|/\delta)\Big).$$

Then, for $K \ge K_1 + K_2$, we have $\mathbb{E}_{s_0\sim d_0}[V_0^{\pi^*}(s_0) - V_0^{\widehat{\pi}}(s_0)] \le \varepsilon$ with probability at least $1-\delta$. So, the total number of samples is at least N_{tot}:

$$N_{\text{tot}} = 2H(K_1 + K_2) = \widetilde{\mathcal{O}}\Big(\frac{\max\{(C(\pi^*))^2, 1\} d H^3 (\lambda + H)^2}{\varepsilon^2} \log^2(H|\mathcal{F}||\mathcal{G}|/\delta)\Big).$$

This completes the proof. □