Distributionally Robust Model-based Reinforcement Learning with Large State Spaces

Shyam Sundhar Ramesh
University College London

Pier Giuseppe Sessa
ETH Zurich

Yifan Hu
EPFL

Andreas Krause
ETH Zurich

Ilija Bogunovic
Univeristy College London

Abstract

Three major challenges in reinforcement learning are the complex dynamical systems with large state spaces, the costly data acquisition processes, and the deviation of real-world dynamics from the training environment deployment. To overcome these issues, we study distributionally robust Markov decision processes with continuous state spaces under the widely used Kullback–Leibler, chi-square, and total variation uncertainty sets. We propose a model-based approach that utilizes Gaussian Processes and the maximum variance reduction algorithm to efficiently learn multi-output nominal transition dynamics, leveraging access to a generative model (i.e., simulator). We further demonstrate the statistical sample complexity of the proposed method for different uncertainty sets. These complexity bounds are independent of the number of states and extend beyond linear dynamics, ensuring the effectiveness of our approach in identifying near-optimal distributionally-robust policies. The proposed method can be further combined with other model-free distributionally robust reinforcement learning methods to obtain a near-optimal robust policy. Experimental results demonstrate the robustness of our algorithm to distributional shifts and its superior performance in terms of the number of samples needed.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1 INTRODUCTION

The use of reinforcement learning (RL) algorithms is gaining momentum in various complex domains, including robotics, nuclear fusion, and molecular discovery. Data acquisition in such environments can be a challenging and resource-intensive process. Safety considerations may also limit the amount of data that can be collected through interactions with the environment. To address this issue, a commonly adopted approach is to train RL policies using a simulator (generative model) enabling RL agents to learn from a simulated environment.

Dealing with complex applications that involve large state spaces requires data-efficient learning, even when a simulator is available. However, achieving optimal policies using existing approaches often requires a significant amount of training data, making data-efficient learning an ongoing challenge. Additionally, when deploying a policy to a real-world system, it is crucial to ensure its performance remains reliable despite mismatches between the simulator and the real-world system. Such mismatches can arise from approximation errors, time-varying system parameters, or even due to adversarial influence. For example, in self-driving, it is infeasible to precisely model all possible variables, such as road conditions, brightness, and tire pressure, which can all vary over time. The resulting mismatch, known as the 'sim-to-real gap', can diminish the performance or impact the reliability of RL algorithms trained on a simulator model.

In this work, we examine the use of a generative model in *distributionally-robust model-based reinforcement learning*. Our aim is to find a distributionally-robust policy that is near-optimal by actively querying the simulator with a state-action pair selected by the learning algorithm. To achieve this, we introduce the kernelized Maximum Variance Reduction (MVR) algorithm, which identifies a state-action pair with the highest un-

certainty according to the model to learn the nominal model dynamics. The algorithm produces a nominal dynamics estimate that is utilized within the robust Markov Decision Process (MDP) framework, where an uncertainty set that includes all models close (according to, e.g., Kullback–Leibler divergence) to the learned one is considered. The overall protocol is summarized in Figure 1. We provide a thorough characterization of statistical sample complexity rates by utilizing the learned model to generate a near-optimal robust policy.

Related Work: Reinforcement learning with a generative model, introduced in Kearns et al. (2002), assumes access to a simulator that outputs the next state given any state-action pair. It appears frequently in the RL literature and is of significant practical relevance. Kakade (2003) elucidate various uses for this generative setting and analyze it in further detail. For the finite MDP case, such a generative setting has been subsequently studied in various works such as Kakade (2003); Gheshlaghi Azar et al. (2013); Li et al. (2020) and, recently, by Agarwal et al. (2020) who provide minimax optimality guarantees for the naive plug-in estimator based algorithm. For large state spaces, generative RL is typically combined with function approximation as studied, e.g., by Abbasi-Yadkori et al. (2019); Shariff and Szepesvári (2020); Lattimore et al. (2020); Li et al. (2023). Recently, Mehta et al. (2021) consider generative RL in continuous state-action spaces from an experimental perspective and showcase the relevance of this setting to the nuclear fusion dynamics research. Degrave et al. (2022) study the tokamak magnetic control problem also using a generative simulator. In addition, Li et al. (2023) present an active exploration strategy that utilizes the least-squares value iteration. Their approach aims to identify a near-optimal policy across the entire state space, providing polynomial sample complexity guarantees that remain unaffected by the number of states. In contrast to these works, we use generative RL to discover *distributionally robust* policies through the modeling of unknown transition dynamics, aiming to address the sim-to-real gap considering the uncertainty in transition dynamics.

The *local access* simulator setting, introduced in Yin et al. (2022), operates under a similar generative model framework. However, the input state to the simulator is restricted to the states already visited. In particular, Tkachuk et al. (2023) study such a setting and employ a model-free approach, focusing on learning the Q-function using an uncertainty-based algorithm. In contrast, our method utilizes the kernel estimator to construct a model of the transition dynamics, offering approximation guarantees. These guarantees are subsequently extended to ensure the robustness of the policy. The novelty of our approach lies in the proposed combination of employing the maximum variance reduction algorithm to learn transition dynamics and constructing a robust policy based on the learned transition dynamics resulting in sample efficiency.

In model-based reinforcement learning, the model learned from a simulator encounters two issues well discussed in the literature, namely, the model-bias (Deisenroth and Rasmussen, 2019; Clavera et al., 2018) and the simulation to reality (sim2real) gap (Andrychowicz et al., 2020; Peng et al., 2018; Mankowitz et al., 2019; Christiano et al., 2016; Rastogi et al., 2018; Wulfmeier et al., 2017). To address this from the perspective of distributional robustness, previous works (Zhou et al., 2021; Panaganti and Kalathil, 2022; Yang et al., 2022) have considered distributional robustness aspects in the context of finite Markov decision processes (MDPs) using the robust MDP framework from Iyengar (2005); Nilim and El Ghaoui (2005). Various other works utilize this robust MDP framework such as Xu and Mannor (2010); Wiesemann et al. (2013); Yu and Xu (2015); Mannor et al. (2016); Badrinath and Kalathil (2021); Petrik and Russel (2019) for the planning problem, and provide asymptotic guarantees for tabular and linear function approximators Lim et al. (2013); Tamar et al. (2014); Roy et al. (2017); Wang and Zou (2021). Our work is closely related to the recent works on distributionally robust RL (Zhou et al., 2021; Panaganti and Kalathil, 2022; Yang et al., 2022; Shi and Chi, 2022; Xu et al., 2023; Clavier et al., 2023; Shi et al., 2024). However, unlike ours, the sample complexity bounds established in these works rely on the number of states and actions, making them impractical for large or infinite state spaces.

A recent work (Blanchet et al., 2024) also consider an infinite state space setup with kernelized function approximation. They consider an offline setting, where the data is already available with the partial coverage assumption satisfied, while we propose to actively collect data from the environment in the generative robust MDPs. Note that the partial coverage assumption introduces an additional variable (coverage coefficient) in their guarantees. Moreover, they propose to utilize the MLE method to estimate their model via offline data. In comparison, we use the maximum variance reduction method to actively generate sample and estimate the model, and demonstrate the sample complexity. Further, their kernel-based transition model is different since they model the transition probability as an inner product between a feature map and a kernel function while we assume that there is a ground truth unknown function. Finally, unlike their work, we consider the χ^2 divergence as well, which invokes a different type of reformulations and analysis. A similar work Ma et al. (2022) deal with linear transition dynamics setup, i.e.,

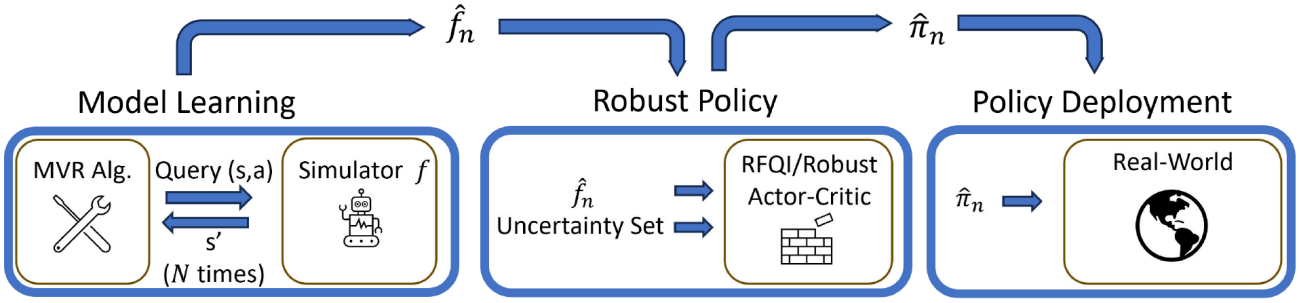


Figure 1: An illustration of distributionally robust RL with a generative model (simulator). Our proposed algorithm MVR (Algorithm 1) queries the simulator and estimates the nominal model $\hat{f}_n$. Then, using the estimated model $\hat{f}_n$ together with a specified uncertainty set, a robust policy is obtained by using model-free RL (e.g., Panaganti et al. (2022)). Finally, the robust policy is deployed in the real-world system.

its Assumption 4.1 is akin to the kernel-based transition dynamics assumption in Blanchet et al. (2024). Still, it also does not involve active sampling.

In the model-free setting with finite state-action space, Liu et al. (2022) propose a distibutionally robust Q-learning algorithm based on access to a generative simulator. Wang et al. (2023a,c) extend the distributionally robust Q-learning framework by improving the design and analysis of the estimation and provide finite sample complexity bounds for this framework. Liang et al. (2023) consider the same problem in the online setting with single trajectory data wherein one is not allowed to sample repeatedly from a state but is allowed to choose actions within that single trajectory. In the model-free setting with large state space (though, still assumed to be finite), Panaganti et al. (2022) study the problem of distributionally robust RL in a function approximation setup. They assume access to offline data from the nominal transition dynamics and provide computational sample complexity bounds in terms of the size of the hypothesis space that is used to represent the set of state-action value functions (Q-function). Other works such as Pinto et al. (2017); Derman et al. (2020); Mankowitz et al. (2019); Zhang et al. (2020) consider robustness aspects in deep reinforcement learning, but these approaches lack theoretical guarantees. To the best of our knowledge, our work is the first one to address the distributionally robust RL problem in the generative model setting with a *model-based* approach and *large* state spaces. Moreover, we are the first to consider general *non-linear* transition dynamics and derive provable sample complexity guarantees for such a setting.

Similar to previous works, we utilize the kernelized MDP framework from Chowdhury and Gopalan (2019) to model transition dynamics with continuous states and actions by assuming that the transition function belongs to an associated Reproducing Kernel Hilbert Space (RKHS). Such continuous MDP formulations also appear in Curi et al. (2020, 2021), however, these works consider finite horizon MDPs while in our work we consider infinite horizon discounted MDPs. In particular, Curi et al. (2021) propose an adversarially robust upper-confidence algorithm to optimize performance in the worst case. However, their algorithm provides robustness guarantees against adversarial perturbations to the transition dynamics. Our work differs from this perspective as we consider robustness w.r.t. distributional shifts of the transition dynamics. Finally, in the related kernelized *bandit* setting, model-based distributionally robust algorithms are proposed in Kirschner et al. (2020); Bogunovic et al. (2018); Nguyen et al. (2020).

We further summarize some recent advancements in distributionally robust reinforcement learning (RL) that have emerged subsequent to our submission. Wang et al. (2023b) focus on developing a comprehensive theoretical framework for robust MDPs by expanding upon existing formulations. Specifically, they explore various types of decision-makers and adversaries' dynamics within the robust MDP framework, including Markovian and history-dependent behaviors, and examine conditions for the existence of the dynamic programming principle. Yang et al. (2023) consider the equivalent Lagrangian version of the robust MDP problem and propose a model-free sample-efficient algorithm to solve the same. Yu et al. (2024) propose a computationally efficient solution for solving robust MDPs with Wasserstein uncertainty set. Li and Shapiro (2023) focus on delineating the connections between static and game formulations of robust MDPs. Unlike our work, the aforementioned works consider the finite state-action space setting. Liu and Xu (2024) study off-dynamics RL through the framework of robust MDPs under total variation uncertainty set and propose a model-free algorithm to learn the robust policy. Panaganti et al. (2023) incorporate techniques from the distributionally robust learning framework to solve the robust MDP problem in the offline RL setting. Both works adopt the linear transition dynamics, differing from our non-linear transition dynamics in the generative model setting with a model-based approach.

Main Contributions: We formalize a distributionally robust reinforcement learning setting with continuous state spaces and non-linear transition dynamics in Section 2. In the generative model setting, we propose (in Section 3) a model-based approach that utilizes Gaussian Process models and the Maximum Variance Reduction (MVR) principle to efficiently learn transition dynamics. We provide novel statistical sample complexity guarantees in Section 4 for the proposed method and widely used uncertainty sets. Our sample complexity bounds are independent of the number of states, ensuring the effectiveness of our approach in identifying near-optimal distributionally robust policies for large state spaces. In Section 5, our experimental findings showcase the sample efficiency and robustness of our algorithm in the face of distributional shifts within popular RL-testing environments.

2 PROBLEM SETTING

A discounted Markov Decision Process (MDP) is a tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, with $\mathcal{S}$ denoting the state space, the action space $\mathcal{A}$, and the probabilistic transition dynamics $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$. Here, $\Delta(\mathcal{S})$ denotes the set of all probability distributions over $\mathcal{S}$. The reward function $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ characterizes the reward $r(s, a)$ the learner receives upon playing $a \in \mathcal{A}$ in $s \in \mathcal{S}$, and $\gamma \in [0, 1]$ denotes the discount factor. The learner uses a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ to select $a \in \mathcal{A}$ upon observing the state $s \in \mathcal{S}$. We define the cumulative discounted reward as $\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$ for known initial state s_0 and $s_t \sim P(s_{t-1}, a_{t-1})$ for $t > 0$ and $a_t \sim \pi(s_t)$. The value function V_π and the state-action value function Q_π are given as follows:

$$V_\pi(s) = \mathbb{E}_{P,\pi}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\Big| s_0 = s\Big],$$

$$Q_\pi(s, a) = r(s, a) + \mathbb{E}_{P,\pi}\Big[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t)\Big],$$

where $a_t \sim \pi(s_t)$ and $s_{t+1} \sim P(s_t, a_t)$. Finally, we define the optimal policy π^* corresponding to dynamics P which yields the optimal value function, i.e., $V_{\pi*}(s) = \max_\pi V_\pi(s)$ for all $s \in \mathcal{S}$.

We assume the standard generative (or random) access model, in which the learner can query transition data arbitrarily from a simulator, i.e., each query to the simulator (s_t, a_t) outputs a sample $s_{t+1} \in \mathbb{R}^d$ where $s_{t+1} \sim P(s_t, a_t)$. In particular, we consider the following frequently used transition dynamics model:

$$s_{t+1} = f(s_t, a_t) + \omega_t, \tag{1}$$

where $\omega_t \in \mathbb{R}^d$ represents independent additive transition noise and follows a Gaussian distribution with zero mean and covariance $\sigma^2 I$.

Regularity assumptions: We assume that f is *unknown* and continuous for tractability reasons which is a common assumption when dealing with continuous state spaces (e.g., Chowdhury and Gopalan (2019); Curi et al. (2020); Kakade et al. (2020)). Further on, we assume that f resides in the Reproducing Kernel Hilbert Space (RKHS). Considering the multi-output definition of f and in line with the previous work (e.g., Chowdhury and Gopalan (2019); Curi et al. (2020)), we define the modified state-action space $\overline{\mathcal{X}}$ (over which the RKHS is defined) as $\overline{\mathcal{X}} := \mathcal{S} \times \mathcal{A} \times [d]$, where the last dimension $i \in \{1, 2, \ldots, d\}$ incorporates the index of the output state vector, i.e., $f(\cdot, \cdot) = (\tilde{f}(\cdot, \cdot, 1), \ldots, \tilde{f}(\cdot, \cdot, d))$ where $\tilde{f} : \overline{\mathcal{X}} \to \mathbb{R}$. In particular, we assume that $\tilde{f}$ belongs to a space of well-behaved functions, denoted by $\mathcal{H}$, induced by some continuous, positive definite kernel function $k : \overline{\mathcal{X}} \times \overline{\mathcal{X}} \to \mathbb{R}$ and equipped with an inner product $\langle \cdot, \cdot \rangle_k$. All functions belonging to an RKHS $\mathcal{H}$ satisfy the reproducing property defined w.r.t. the inner product $\langle \cdot, \cdot \rangle_k : \langle \tilde{f}, k(x, \cdot) \rangle = \tilde{f}(x)$ for $\tilde{f} \in \mathcal{H}$. We also make the following common assumptions: (i) the kernel function k is bounded $k(x, x') \leqslant 1$ for all $x, x' \in \overline{\mathcal{X}}$ and $\overline{\mathcal{X}}$ is a compact set ($\mathcal{X} \subset \mathbb{R}^p$), and (ii) every function $\tilde{f} \in \mathcal{H}$ has a bounded RKHS norm (induced by the inner product) $\|\tilde{f}\|_k \leqslant B$.

We refer to the simulator environment determined by f as the *nominal model* P_f, while the true environment encountered by the agent in the real world might not be the same (e.g., due to a sim-to-real gap). Consequently, we utilize the robust MDP framework to tackle this by considering an uncertainty set comprising of all models close to the nominal one.

Robust Markov Decision Process (RMDP): We consider the robust MDP setting that addresses the uncertainty in transition dynamics and considers a set of transition models called the *uncertainty set*. We use $\mathcal{P}^f$ to denote the uncertainty set that satisfies the (s, a)–rectangularity condition Iyengar (2005) (as defined in Equation (2)), an assumption that is commonly used in the related literature (e.g., Panaganti and Kalathil (2022); Panaganti et al. (2022); Zhou et al. (2021)). Similar to MDPs, we specify RMDP by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}^f, r, \gamma)$ where the uncertainty set $\mathcal{P}^f$ consists of all models close to a nominal model P_f in terms of a distance measure D:

$$\mathcal{P}^f_{s,a} = \{p \in \Delta(\mathcal{S}) : D(p||P_f(s, a)) \leqslant \rho\},$$
$$\mathcal{P}^f = \bigotimes_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mathcal{P}^f_{s,a}. \tag{2}$$

Here, D denotes some distance measure between probability distributions, and $\rho > 0$ defines the radius of the uncertainty set. In this work, we consider three probability distance measures, including Kullback–Leibler (KL) divergence such that $\mathrm{KL}(P||Q) = \int \log(\frac{dP}{dQ}) dP$, Chi-Square ($\chi^2$) distance

such that $\chi^2(P||Q) = \int(\frac{dP}{dQ} - 1)^2 dP$ for P being absolutely continuous with respect to Q, and Total Variation (TV) distance such that $\mathrm{TV}(P||Q) = \frac{1}{2}\|P - Q\|_1$.

In the RMDP setting, the goal is to discover a policy that maximizes the cumulative discounted reward for the worst-case transition model within the *given* uncertainty set. Concretely, the robust value function $V^R_{\pi,f}$ corresponding to a policy π and the optimal robust value function are given as follows:

$$V^{\mathrm{R}}_{\pi,f}(s) = \inf_{P\in\mathcal{P}^f} \mathbb{E}_{P,\pi}\Big[\sum_{t=1}^{\infty}\gamma^t r(s_t,a_t)\Big|s_0 = s\Big],$$
$$V^{\mathrm{R}}_{\pi*,f}(s) = \max_{\pi} V^{\mathrm{R}}_{\pi,f}(s) \quad \forall s\in\mathcal{S}. \tag{3}$$

In fact, Iyengar (2005) shows that for any f, there exists a deterministic robust policy π^*. Using the definition of the robust value function, we also define the robust Bellman operator (Iyengar, 2005) in terms of the robust state-action value function $Q^R_{\pi,f}$ as follows:

$$Q^{\mathrm{R}}_{\pi,f}(s,a) = r(s,a) + \gamma \inf_{D(p||P_f(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\pi,f}(s')\Big]. \tag{4}$$

The goal of the learner is to discover a near-optimal robust policy while minimizing the total number of samples N, i.e., queries to the nominal model (simulator). Concretely, for a fixed precision $\epsilon > 0$, the goal is to output a policy $\hat{\pi}_N$ after collecting N samples, such that $\|V^{\mathrm{R}}_{\hat{\pi}_N,f} - V^{\mathrm{R}}_{\pi*,f}\|_\infty \leqslant \epsilon$.

3 SAMPLING ALGORITHM

In this section, we outline our methodology for addressing the problem described in Section 2. We begin by discussing the Gaussian process (GP) model often used in algorithms to learn RKHS functions (Rasmussen and Williams, 2006; Kanagawa et al., 2018).

Multi-output Gaussian process: Under the assumptions of Section 2, modeling uncertainty and learning the transition model f can be performed via the Gaussian process framework. A Gaussian process $GP(\mu(\cdot), k(\cdot,\cdot))$ over the input domain $\overline{\mathcal{X}}$, is a collection of random variables $(\tilde{f}(x))_{x\in\overline{\mathcal{X}}}$ whose every finite subset $(\tilde{f}(x_i))_{i=1}^n, n\in\mathbb{N}$, follows multivariate Gaussian distribution with mean $\mathbb{E}[\tilde{f}(x_i)] = \mu(x_i)$ and covariance $\mathbb{E}[(\tilde{f}(x_i) - \mu(x_i))(\tilde{f}(x_j) - \mu(x_j))] = k(x_i, x_j)$ for every $1 \leqslant i, j \leqslant n$. Standard algorithms implicitly use a zero-mean $GP(0, k(\cdot,\cdot))$ as the prior distribution over $\tilde{f}$, i.e, $\tilde{f} \sim GP(0, k(\cdot,\cdot))$, and assume that the noise variables are drawn independently across t from $\mathcal{N}(0,\lambda)$ with $\lambda > 0$. Considering the multi-output definition of $f(\cdot,\cdot) = (\tilde{f}(\cdot,\cdot,1),\ldots,\tilde{f}(\cdot,\cdot,d))$,we build d copies of the dataset such that $\mathcal{D}_{1:n,l} = \{(s_i,a_i,l), s_{i+1,l}\}_{i=1}^n$ each with n transitions from a particular state-action pair (s,a) to component l of next state. For $x_i = (s_i,a_i)$ and $y_{i,l} = s_{i+1,l}$, the posterior mean, covariance and variance for $\tilde{f}(x,l)$ are given by:

$$\mu_{nd}(x,l) = k_{nd}(x,l)(K_{nd} + I_{nd}\lambda)^{-1} y_{nd}, \tag{5}$$
$$k_{nd}((x,l),(x',l)) = k((x,l),(x',l)) - k_{nd}(x,l)(K_{nd} + I_{nd}\lambda)^{-1} k^T_{nd}(x',l'),$$
$$\sigma^2_{nd}(x,l) = k_{nd}((x,l),(x,l)). \tag{6}$$

Here K_{nd} denotes the kernel matrix of dimensions $nd \times nd$ whose entries are $k((x_i,l),(x_j,l'))$ with $1 \leqslant i,j \leqslant n$ and $1 \leqslant l,l' \leqslant d$. $k_{nd}(x,l) = [k((x,l),(x_i,l'))]_{1\leqslant i\leqslant n, 1\leqslant l'\leqslant d}$ denotes the covariance vector and $y_{nd} = [y_{i,l}]_{1\leqslant i\leqslant n, 1\leqslant l\leqslant d}$ denotes the output vector.

Correspondingly, the posterior mean and variance for f would be

$$\mu_n(s,a) = (\mu_{nd}(s,a,1),\cdots,\mu_{nd}(s,a,d)), \tag{7}$$
$$\sigma_n(s,a) = (\sigma_{nd}(s,a,1),\cdots,\sigma_{nd}(s,a,d)). \tag{8}$$

Maximum variance reduction: With certain assumptions on the loss function (squared loss) and noise distribution, the function estimation in RKHS is analogous to the Bayesian Gaussian process framework (Rasmussen and Williams, 2006). When used with the same kernel function, this allows the construction of mean and variance estimates of $\tilde{f}\in\mathcal{H}$ using Gaussian processes (eq. (5) and eq. (6)). Based on these, one can construct shrinking statistical confidence bounds (used in our analysis in Appendix A.2) that hold with probability at least $1-\delta$, i.e., the following holds $|\tilde{f}(x) - \mu_{n-1}(x)| \leqslant \beta_n(\delta)\sigma_{n-1}(x)$ for every $n \geqslant 1$ and $x\in\overline{\mathcal{X}}$. Here $\{\beta_i\}_{i=1}^n$ stands for the sequence of parameters that are suitably set (see Lemma 5) to ensure the validity of the confidence bounds.

We use the maximum variance reduction (MVR) algorithm (Algorithm 1) to learn about the nominal model f. MVR works on the principle of reducing the maximum uncertainty measured by the posterior standard deviation of a GP model calculated by using previously collected data. At each iteration, MVR queries a state-action pair that has the highest uncertainty according to the model and uses the obtained observation to update the GP posterior. The algorithm outputs nominal dynamics estimate $\hat{f}_n$ corresponding to the final GP posterior mean μ_n.

To characterize the precision of the learned model, we use the max. information gain (Srinivas et al., 2009)

$$\Gamma_n(\overline{\mathcal{X}}) = \max_{x_1,\ldots,x_n\in\overline{\mathcal{X}}} 0.5\log\det(I_n + \lambda^{-1}K_n), \tag{9}$$

a kernel-dependent quantity that is frequently used in GP optimization. For many commonly used kernels, Γ_n is sublinear in n, which implies that the predictive

Algorithm 1 Maximum Variance Reduction (MVR) for learning model dynamics

Require: Simulator f, kernel k, domain $\mathcal{S} \times \mathcal{A}$
Set $\mu_0(s,a) = 0$, $\sigma_0(s,a) = 1$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$
for $i = 1, \ldots, n$ **do**
  $(s_i, a_i) = \arg\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\sigma_{i-1}(s,a)\|_2$
  Observe $s_{i+1} = f(s_i, a_i) + \omega_i$ (i.e., sample s_{i+1} from nominal $P_f(s_i, a_i)$)
  Update to μ_i and σ_i by using (s_i, a_i, s_{i+1}) according to Eq. (5) and Eq. (6)
end for
return The dynamics estimate $\hat{f}_n(\cdot,\cdot) = \mu_n(\cdot,\cdot)$

uncertainties are shrinking sufficiently fast, and thus $\hat{f}_n$ is capable of generalizing well across the entire domain. This is formalized in the following lemma.

Lemma 1. *For $\beta_n(\delta)$ set as in Lemma 4 and $\mathcal{I}_d$ denoting $\{1, 2, \cdots, d\}$, the MVR algorithm (Algorithm 1) outputs the dynamics estimate $\hat{f}_n(\cdot,\cdot) = \mu_n(\cdot,\cdot)$ such that the following holds uniformly for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ with probability at least $1-\delta$,*

$$\|\mu_n(s,a) - f(s,a)\|_2 \leqslant \mathcal{O}\Big(\frac{\beta_n(\delta) 2ed}{\sqrt{n}} \sqrt{\Gamma_{nd}(\mathcal{S} \times \mathcal{A} \times \mathcal{I}_d)}\Big).$$

The preceding lemma asserts that we can effectively estimate the unknown dynamics by utilizing the pure exploration procedure and that the error in the model reduces as we increase the number of samples. In the subsequent section, we leverage this finding to establish the minimum number of samples needed to obtain a distributionally robust policy that is close to optimal.

4 SAMPLE COMPLEXITY

This section discusses the statistical sample complexity of the proposed MVR algorithm in distributionally robust MDPs. Specifically, given the optimal robust policies $\hat{\pi}_N$ and π^* corresponding to the learned nominal dynamics $\hat{f}_N$ by the MVR algorithm with N iterations and the true nominal dynamics f, respectively, we show the number of samples needed by the MVR algorithm to ensure that the following holds:

$$|V^{\mathrm{R}}_{\hat{\pi}_N, f}(s) - V^{\mathrm{R}}_{\pi^*, f}(s)| \leqslant \epsilon, \forall\ s \in \mathcal{S}. \tag{10}$$

Note that such an assumption on access to optimal policy is widely used in the generalization literature (Kleywegt et al., 2002; Hu et al., 2020; Zhang et al., 2024). Several model-free methods (Panaganti et al., 2022; Derman et al., 2018; Mankowitz et al., 2019) have studied how to learn an optimal robust policy under KL, TV, and χ^2 uncertainty set given trajectory samples generated from a transition dynamics. Thus, one can easily incorporate the MVR algorithm with these model-free algorithms to find an optimal $\hat{\pi}_N$ using samples generated by $\hat{f}_N$. One major benefit of our approach is that we do not need access to samples from the more costly simulator f when training the robust policy. Consequently, we focus on the statistical sample complexity of the MVR algorithm rather than designing algorithms to find $\hat{\pi}_N$.

Theorem 1. *(Sample Complexity of MVR under KL uncertainty set) Consider a robust MDP with nominal transition dynamics f satisfying the regularity assumptions from Section 2 and with uncertainty set defined as in Equation (2) w.r.t. KL divergence. For π^* denoting the robust optimal policy w.r.t. nominal transition dynamics f and $\hat{\pi}_N$ denoting the robust optimal policy w.r.t. learned nominal transition dynamics $\hat{f}_N$ via MVR (Algorithm 1), and $\delta \in (0,1)$, $\epsilon \in (0, \frac{1}{1-\gamma})$, it holds that $\max_s |V^R_{\hat{\pi}_N, f}(s) - V^R_{\pi^*, f}(s)| \leqslant \epsilon$ with probability at least $1-\delta$ for any N such that*

$$N = \mathcal{O}\Big(e^{\frac{2-\gamma}{(1-\gamma)\alpha_{\mathrm{kl}}}} \frac{\gamma^2 \beta_N^2(\delta) d^2 \Gamma_{Nd}}{(1-\gamma)^4 \rho^2 \epsilon^2}\Big). \tag{11}$$

Theorem 1 shows the number of samples required from the nominal transition dynamics f (simulator) to construct a robust optimal policy reliably with high probability. The complexity bound depends on the maximum information gain Γ_{Nd} (Equation (9)), which is sublinear in N for many commonly used kernels (Srinivas et al., 2009), and on the confidence width $\beta_N^2(\delta)$ which also exhibits a logarithmic dependence on N according to the confidence bounds from Vakili et al. (2021). Specifically, for the squared exponential kernel used in the experiments, both Γ_{Nd} and β_N depend only logarithmically on N, implying that a bound independent of N remains the same up to extra multiplicative log factors. An additional d factor that denotes the dimension of the state space $\mathcal{S}$ in the obtained bound comes from utilizing the multi-output (of dimension d) GP framework to model the transition dynamics, which also appears in the regret bounds of similar works (Chowdhury and Gopalan, 2019; Curi et al., 2020, 2021). Finally, the term $\alpha_{\mathrm{kl}} \in (0, \frac{1}{2(1-\gamma)\rho})$ is a problem-dependent parameter that is independent of N, which similarly appears in the guarantees of Panaganti and Kalathil (2022).

We can compare our guarantees with the existing sample-complexity results in model-based distributionally robust RL which, however, only consider finite state-action spaces. In particular, when considering KL uncertainty sets with infinite horizon, Panaganti and Kalathil (2022) obtain sample complexity of order $\mathcal{O}\Big(e^{\frac{\alpha_{\mathrm{kl}}+2}{\alpha_{\mathrm{kl}}(1-\gamma)}} \frac{\gamma^2 |\mathcal{S}|^2 |\mathcal{A}|}{(1-\gamma)^4 \rho^2 \epsilon^2}\Big)$ up to logarithmic factors. Notably, the latter complexity bound explicitly depends on the cardinality of the state and action spaces $|\mathcal{S}|$ and $|\mathcal{A}|$, thus scaling badly when $\mathcal{S}$ and $\mathcal{A}$ are large or continuous. Instead, the guarantee of Theorem 1 depends on the state-action space *only* through Γ_{Nd} which remains bounded even when these are continuous. This

allows us to successfully extend the distributionally robust framework to continuous state spaces. Other terms in the bound of Theorem 1 such as γ (the discount factor), ρ (radius of the uncertainty set) have similar dependencies. Crucially, the dependence on the precision parameter ϵ remains the same when compared to the guarantees provided for finite state-action setting.

We relegate the proof of Theorem 1 to Appendix B but outline its main steps below:

Step (i): The first step is to bound the approximation error of policy $\hat{\pi}_n$ (i.e., the left-hand side of Equation (10)) by the sum of two error terms: $|V^{\mathrm{R}}_{\hat{\pi}_N,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_N,\hat{f}_N}(s)|$ and $|V^{\mathrm{R}}_{\hat{\pi}_N,\hat{f}_N}(s) - V^{\mathrm{R}}_{\pi^*,f}|$. Utilizing the robust Bellman Equation (4), bounding such errors boils down to bounding differences of the form:

$$\max_s \Big| \inf_{\mathrm{KL}(p||P_f(s,\hat{\pi}_N(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_N,f}(s')\Big] - \inf_{\mathrm{KL}(p||P_{\hat{f}_n}(s,\hat{\pi}_N(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_N,f}(s')\Big]\Big|. \quad (12)$$

where $P_f(s,a)$ denotes the Gaussian transition distribution with mean $f(s,a)$ and covariance $\sigma^2 I$.

Step (ii): The major challenge of bounding Equation (12) lies in the inner infinite-dimensional minimization problems over distributions. To overcome this, we can reformulate such problems into single-dimensional ones using duality (Hu and Hong, 2013; Zhou et al., 2021; Panaganti and Kalathil, 2022) as follows:

Lemma 2. *(Variant of Hu and Hong (2013)) For random variable X and function V satisfying that $V(X)$ has a finite Moment Generating function, it holds for all $\rho > 0$ and $P \in \{P : \mathrm{KL}(P||P_0) \leqslant \rho\}$:*

$$\inf_P \mathbb{E}_P[V(X)] = \sup_{\alpha\geqslant 0}\{-\alpha \log(\mathbb{E}_{P_0}[e^{\frac{-V(X)}{\alpha}}]) - \alpha\rho\}.$$

Let $H(V,P) := \sup_{\alpha\geqslant 0}\{-\alpha \log(\mathbb{E}_{X\sim P}[e^{\frac{-V(X)}{\alpha}}]) - \alpha\rho\}$. Thus, applying Lemma 2, we rewrite Equation (12) as the difference of two single-dimensional convex optimization problems with expectations over P_f and $P_{\hat{f}_N}$, respectively:

$$\begin{aligned}
&\max_s \Big| H(V^{\mathrm{R}}_{\hat{\pi}_N,f}, P_f(s,\hat{\pi}_N(s))) - H(V^{\mathrm{R}}_{\hat{\pi}_N,f}, P_{\hat{f}_N}(s,\hat{\pi}_N(s)))\Big| \\
&\leqslant \max_{V(\cdot)\in\mathcal{V}} \max_{s,a} \Big| H(V, P_f(s,a)) - H(V, P_{\hat{f}_N}(s,a))\Big| \\
&\leqslant \max_{V(\cdot)\in\mathcal{V}} \max_{s,a} \max_{\alpha\in[\underline{\alpha},\overline{\alpha}]} c\Big|\mathbb{E}_{s'\sim P_{f(s,a)}}[e^{\frac{-V(s')}{\alpha}}] \\
&\qquad\qquad - \mathbb{E}_{s'\sim P_{\hat{f}_N(s,a)}}[e^{\frac{-V(s')}{\alpha}}]\Big|,
\end{aligned} \quad (13)$$

where c, $\underline{\alpha}$, $\overline{\alpha} > 0$ are constants, $\mathcal{V}$ denotes the value functional space, and the last inequality holds due to certain structural properties of the single-dimensional optimization problem in the RHS of Equation (13).

Step (iii): Finally, we bound Equation (13) using the difference between the estimated model $\hat{f}_N$ and the true f, which is characterized by Lemma 1, in Appendix B. Moreover, to address the outer maximum over all value functions, states, and actions, we incorporate a covering number argument.

Other uncertainty sets: We further obtain the statistical sample complexities for χ^2 distance and TV distance uncertainty sets. We note that the analysis follows similar steps as the ones of Theorem 1. The major difference lies in incorporating and handling the dual forms of χ^2/TV uncertainty sets in our analysis which differ from the one of Lemma 2. For χ^2 uncertainty set, we utilize the dual formulation that appears in Duchi and Namkoong (2021), while for TV uncertainty sets we follow the approach of Yang et al. (2022). As before, we can upper bound Equation (12) via covering number arguments and the distance between the nominal transition dynamics f and the learned transition dynamics $\hat{f}_N$ by using Lemma 1. Below, we outline the statistical sample complexity in the case of χ^2 and TV uncertainty sets in Propositions 1 and 2, respectively.

Proposition 1. *(Sample Complexity of MVR under χ^2 uncertainty set) Under the setup of Theorem 1 with uncertainty set defined w.r.t. χ^2 distance, it holds that $\max_s |V^R_{\hat{\pi}_N,f}(s) - V^R_{\pi^*,f}(s)| \leqslant \epsilon$ with probability at least $1-\delta$ for any N such that*

$$N = \mathcal{O}\Big(\Big(\tfrac{1+2\rho}{\sqrt{1+2\rho}-1}\Big)^4 \tfrac{\gamma^4\beta^2_N(\delta)d^2\Gamma_{Nd}}{(1-\gamma)^8\epsilon^4}\Big). \quad (14)$$

Proposition 2. *(Sample Complexity of MVR under TV uncertainty set) Under the setup of Theorem 1 with uncertainty set defined w.r.t. TV distance, it holds that $\max_s |V^R_{\hat{\pi}_N,f}(s) - V^R_{\pi^*,f}(s)| \leqslant \epsilon$ with probability at least $1-\delta$ for any N such that*

$$N = \mathcal{O}\Big(\tfrac{(2+\rho)^2}{\rho^2}\tfrac{\gamma^2\beta^2_N(\delta)d^2\Gamma_{Nd}}{(1-\gamma)^4\epsilon^2}\Big). \quad (15)$$

We relegate the proofs of Propositions 1 and 2 to Appendices C.1 and C.2. In comparison to the exponential dependence on $\frac{1}{1-\gamma}$ for KL uncertainty set in Theorem 1, we note that for both χ^2/TV uncertainty sets, we obtain *polynomial* dependence on $\frac{1}{1-\gamma}$. In the context of the TV uncertainty set, the dependency on ϵ in Proposition 2 remains consistent with the finite state case (Panaganti and Kalathil, 2022). However, in the χ^2 case, the bound presented in Proposition 1 exhibits a worse dependence on ϵ compared to the result derived in Panaganti and Kalathil (2022). This difference arises because we refrain from utilizing the same dual reformulation lemmas from Iyengar (2005), as they are applicable exclusively to finite state-action settings. Improving these rates is an interesting direction for future work.

5 EXPERIMENTS

The aim of our experiments is to show the effectiveness of the proposed distributionally-robust model-based approach. In particular, our goal is to evaluate the robustness of our policies against different perturbations of the environment's parameters, and compare them with existing non-robust methods. Moreover, our experiments focus on demonstrating the effectiveness of MVR to smartly collect data from the environment rather than using a sub-optimal/random policy to interact with the environment and collect data. This significantly reduces the number of samples required from the environment to perform robustly. In addition, we compare our approach with model-free methods (robust and non-robust) which typically require a significantly larger number of interactions with the nominal environment.

Environments: We consider the OpenAI's gym (Brockman et al., 2016) environments of swing-up Pendulum, Cartpole and Reacher, respectively. Pendulum has a 2-dimensional state space and scalar actions (Mehta et al., 2021). For Cartpole, we consider a scalar continuous action space as done in Mehta et al. (2021), while states are 4-dimensional. Reacher, instead, consists of a 2DOF robot arm with 8-dimensional states. For each environment we test our approach against various perturbations as outlined below.

Module 1: Learning the model. To learn the nominal environment, we utilize a setup similar to that of Mehta et al. (2021), but instead of considering the "EIG$_{\tau*}$" which minimizes entropy of the optimal trajectory τ^* using model-predictive control, we use the proposed MVR method (Algorithm 1). Similar to Mehta et al. (2021), we use a GP prior with the squared exponential kernel to model the transition dynamics $f(s, a)$ (alternate models such as Neural Ensembles or Bayesian neural networks can be used to model the transition dynamics as done in, e.g., Curi et al. (2020, 2021)). As in continuous control problems the subsequent states are fairly close, we use our multi-output GP to model the difference $f(s_t, a_t) - s_{t+1}$.

Module 2: Computing a robust policy. Given a learned model $\hat{f}_n$, we compute the associated robust policy $\hat{\pi}_n$ using the Robust Fitted Q-Iteration (RFQI) algorithm recently introduced in Panaganti et al. (2022) (this effectively approximates our robust optimization oracle). RFQI computes a robust policy from offline data by alternated maximization of a dual-variable function and a Q-function. We generate such offline data by using a ϵ-greedy non-robust policy (using Soft Actor-Critic (Haarnoja et al., 2018) or Model Predictive Control (Camacho and Alba, 2013; Chua et al., 2018)) which we train *on the learned model* $\hat{f}_n$ from Module 1 and let interact with it for 10^6 steps. Note that this is crucially different from the vanilla RFQI (Panaganti et al., 2022) where the true nominal environment was used both for training such policy and for generating offline data. Indeed, this would require a significantly larger number of environment interactions.

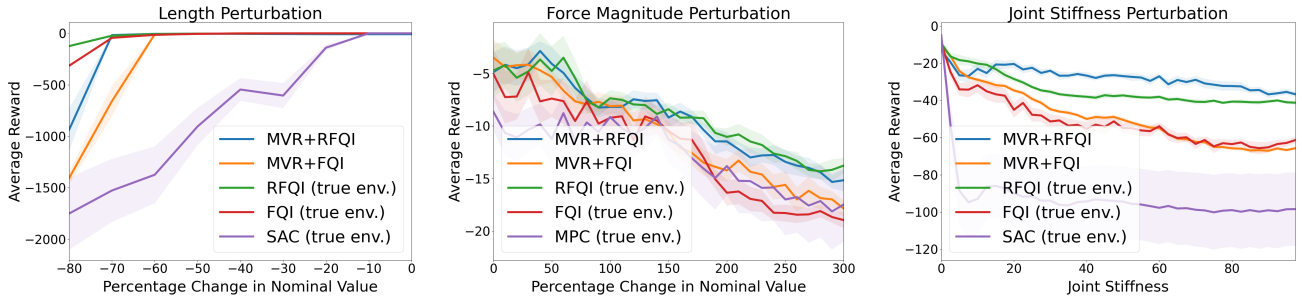
Alg \ Env	Pendulum	Cartpole	Reacher
MVR+RFQI (ours)	**60**	**150**	**2000**
MVR+FQI	60	150	2000
SAC	10^4	-	10^6
MPC	-	2250/step	-
RFQI	$10^6 + 10^4$	$10^5 \cdot 2250$	$10^6 + 10^6$
FQI	$10^6 + 10^4$	$10^5 \cdot 2250$	$10^6 + 10^6$

Table 1: Number of interactions with the nominal environment to obtain the results of Figure 2. For MPC, a total of 2250 interactions are required at each step for planning multiple rollouts and selecting the best action. Both RFQI and FQI utilize 10^6 offline data points generated by SAC or MPC.

Baselines: We compare our approach, which we denote as MVR+RFQI, with the following baselines:

- MVR+FQI: A natural non-robust baseline that consists of computing a non-robust policy via the Fitted Q-Iteration (FQI) algorithm (Ernst et al., 2005) on the same offline data used by MVR+RFQI,
- Soft Actor-Critic (SAC) (Haarnoja et al., 2018), or Model Predictive Control (MPC) (Camacho and Alba, 2013; Chua et al., 2018), as model-free methods which compute non-robust policies *interacting with the nominal environment* (in case of MPC, the latter is used for planning),
- RFQI (Panaganti et al., 2022), which also requires the nominal environment and uses 10^6 offline data collected by SAC or MPC to train a robust policy,
- FQI (Ernst et al., 2005), which trains a non-robust policy from the same data.

Training: Model-free methods are trained directly on the nominal environments. In particular, for Pendulum and Reacher we train SAC until convergence for 10^4 and 10^6 steps, respectively. On the continuous actions Cartpole, instead, we run MPC following the implementation of Pinneri et al. (2020); Mehta et al. (2021) which requires a total of 2250 planning interactions to select the optimal action at each step. Depending on the environment, we utilize SAC or MPC mixed with an ϵ-greedy rule to collect 10^6 offline data points. These are used to train the offline methods RFQI and FQI as done in Panaganti et al. (2022). For the model-based approaches, instead, we first run MVR for a sufficiently informative number of samples (60 for Pendulum, 150 for Cartpole and 2000 for Reacher) to obtain an estimated model $\hat{f}_n$. Then, we use SAC (trained against model $\hat{f}_n$) or MPC to collect 10^6 offline data on such estimated environment.



(a) Pendulum (b) Cartpole (c) Reacher

Figure 2: Average performance (over 20 episodes) on the considered environments, as a function of different perturbations: length perturbation for Pendulum, force magnitude perturbation for Cartpole, and perturbed joint stiffness for Reacher. Unlike our MVR+RFQI and non-robust MVR+FQI, the other baselines are model-free and require access to the true nominal environment for training. The proposed approach MVR+RFQI achieves comparable performance to the model-free RFQI albeit requiring significantly fewer environment interactions (see Table 1). Moreover, as the perturbation magnitude increases, MVR+RFQI outperforms the other non-robust baselines.

These data are then used to train MVR+RFQI and MVR+FQI. We provide further implementation details and hyperparameters in Appendix D.

Evaluation: For each environment, we evaluate the computed policy against different perturbation types and magnitudes. For Cartpole, we perturb the magnitude of the actuation force. Its nominal value is 10 and we perturb up to 300%. Also, we consider perturbations to gravity in the range of (-100%,100%) with the nominal value being 9.82. For the Pendulum, we consider perturbations to the length of the pendulum and action perturbations (where a random action is chosen with ϵ probability). Finally, in the case of Reacher we consider perturbations to the joint's stiffness (from 0 to 100) coupled with perturbations of the joint's equilibrium position. Further details on the chosen perturbations and hyperparameters used are provided in Appendix D. We provide the code to reproduce the results.[1]

In Figure 2 we plot the average performance (over 20 episodes) of the baselines subject to different perturbation types and magnitudes for each environment. Results for other perturbations are relegated to Appendix D. In Table 1, we report the total number of interactions with the nominal environment required to compute the evaluated policies. We remark that MVR+RFQI and MVR+FQI interact with the environment only to learn a GP model via the MVR approach. Instead, the other model-free methods use the nominal environment throughout the whole training and, in case of RFQI and FQI, even to generate offline data. Notably, the policy computed by MVR+RFQI displays comparable performance to its model-free counterpart RFQI which, as shown in Table 1, requires a significantly larger number of samples. This shows the sample-efficiency of MVR in acquiring informative data. Moreover, as the perturbation magnitude increases, MVR+RFQI achieves higher performance compared to MVR+FQI and the other non-robust methods, demonstrating the robustness of the computed policies. Additionally, as similarly noted e.g. by Kumar et al. (2021), we observe the offline methods MVR+FQI and FQI to be generally more robust (although not explicitly computing robust policies) than SAC and MPC.

[1] https://github.com/rsshyam/MVR-RFQI

6 CONCLUSIONS

We investigated distributionally robust RL with continuous state spaces and non-linear transitions. Specifically, we proposed a model-based approach in the generative model setting, utilizing max. variance reduction to learn nominal transitions. Our results include novel sample complexity guarantees for commonly used uncertainty sets, required for identifying near-optimal robust policies in large state spaces. Through experiments conducted in popular RL-environments, we demonstrated the sample efficiency and robustness of our algorithm in the presence of distributional shifts. An important avenue is the extension of our algorithm to the online and offline RL settings.

Acknowledgements

PGS was gratefully supported by ELSA (European Lighthouse on Secure and Safe AI) funded by the European Union under grant agreement No. 101070617. YH was supported by NCCR Automation from Switzerland. IB was supported by the EPSRC New Investigator Award EP/X03917X/1 and Google Research Scholar award. The authors would like to thank Viraj Mehta, Zaiyan Xu, Zhengqing Zhou, Zhengyuan Zhou, Wenhao Yang, Laixi Shi, and Liangyu Zhang for the useful discussion.

References

Abbasi-Yadkori, Y., Bartlett, P., Bhatia, K., Lazic, N., Szepesvari, C., and Weisz, G. (2019). Politex: Regret bounds for policy iteration using expert prediction. *International Conference on Machine Learning (ICML)*.

Agarwal, A., Kakade, S., and Yang, L. F. (2020). Model-based reinforcement learning with a generative model is minimax optimal. *Conference on Learning Theory (COLT)*.

Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*.

Andrychowicz, O. M., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., et al. (2020). Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*.

Badrinath, K. P. and Kalathil, D. (2021). Robust reinforcement learning using least squares policy iteration with provable performance guarantees. *International Conference on Machine Learning (ICML)*.

Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G. (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Management Science (INFORMS)*.

Blanchet, J., Lu, M., Zhang, T., and Zhong, H. (2024). Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *Conference on Neural Information Processing Systems (NeurIPS)*.

Bogunovic, I., Scarlett, J., Jegelka, S., and Cevher, V. (2018). Adversarially robust optimization with Gaussian processes. *Conference on Neural Information Processing Systems (NeurIPS)*.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). Openai gym.

Camacho, E. F. and Alba, C. B. (2013). *Model predictive control*. Springer Science & Business media.

Chen, J. and Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. *International Conference on Machine Learning (ICML)*.

Chowdhury, S. R. and Gopalan, A. (2019). Online learning in kernelized Markov decision processes. *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Christiano, P., Shah, Z., Mordatch, I., Schneider, J., Blackwell, T., Tobin, J., Abbeel, P., and Zaremba, W. (2016). Transfer from simulation to real world through learning deep inverse dynamics model. *arXiv preprint arXiv:1610.03518*.

Chua, K., Calandra, R., McAllister, R., and Levine, S. (2018). Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Conference on Neural Information Processing Systems (NeurIPS)*.

Clavera, I., Rothfuss, J., Schulman, J., Fujita, Y., Asfour, T., and Abbeel, P. (2018). Model-based reinforcement learning via meta-policy optimization. *Conference on Robot Learning*.

Clavier, P., Pennec, E. L., and Geist, M. (2023). Towards minimax optimality of model-based robust reinforcement learning. *arXiv preprint arXiv:2302.05372*.

Cressie, N. and Read, T. R. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B (Methodological)*.

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

Curi, S., Berkenkamp, F., and Krause, A. (2020). Efficient model-based reinforcement learning through optimistic policy search and planning. *Conference on Neural Information Processing Systems (NeurIPS)*.

Curi, S., Bogunovic, I., and Krause, A. (2021). Combining pessimism with optimism for robust and efficient model-based deep reinforcement learning. *International Conference on Machine Learning (ICML)*.

Degrave, J., Felici, F., Buchli, J., Neunert, M., Tracey, B., Carpanese, F., Ewalds, T., Hafner, R., Abdolmaleki, A., de Las Casas, D., et al. (2022). Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*.

Deisenroth, M. P. and Rasmussen, C. E. (2019). Pilco: A model-based and data-efficient approach to policy search. *International Conference on Machine Learning (ICML)*.

Derman, E., Mankowitz, D., Mann, T., and Mannor, S. (2020). A Bayesian approach to robust reinforcement learning. *Uncertainty in Artificial Intelligence (UAI)*.

Derman, E., Mankowitz, D. J., Mann, T. A., and Mannor, S. (2018). Soft-robust actor-critic policy-gradient. *arXiv preprint arXiv:1803.04848*.

Duchi, J. C. and Namkoong, H. (2021). Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics, Institute of Mathematical Statistics*.

Ernst, D., Geurts, P., and Wehenkel, L. (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research (JMLR)*.

Gheshlaghi Azar, M., Munos, R., and Kappen, H. J. (2013). Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Springer Machine Learning*.

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

Hu, Y., Chen, X., and He, N. (2020). Sample complexity of sample average approximation for conditional stochastic optimization. *SIAM Journal on Optimization*.

Hu, Z. and Hong, L. J. (2013). Kullback-Leibler divergence constrained distributionally robust optimization. *Optimization Online*.

Iyengar, G. N. (2005). Robust dynamic programming. *Mathematics of Operations Research (INFORMS)*.

Kakade, S., Krishnamurthy, A., Lowrey, K., Ohnishi, M., and Sun, W. (2020). Information theoretic regret bounds for online nonlinear control. *Conference on Neural Information Processing Systems (NeurIPS)*.

Kakade, S. M. (2003). *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom).

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. (2018). Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*.

Kearns, M., Mansour, Y., and Ng, A. Y. (2002). A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Springer Machine learning*.

Kirschner, J., Bogunovic, I., Jegelka, S., and Krause, A. (2020). Distributionally robust Bayesian optimization. *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Kleywegt, A. J., Shapiro, A., and Homem-de Mello, T. (2002). The sample average approximation method for stochastic discrete optimization. *SIAM Journal on optimization*.

Kumar, A., Hong, J., Singh, A., and Levine, S. (2021). Should i run offline reinforcement learning or behavioral cloning? *International Conference on Learning Representations (ICLR)*.

Lattimore, T., Szepesvari, C., and Weisz, G. (2020). Learning with good feature representations in bandits and in rl with a generative model. *International Conference on Machine Learning (ICML)*.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2020). Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Conference on Neural Information Processing Systems (NeurIPS)*.

Li, X., Mehta, V., Kirschner, J., Char, I., Neiswanger, W., Schneider, J., Krause, A., and Bogunovic, I. (2023). Near-optimal policy identification in active reinforcement learning. *International Conference on Learning Representations (ICLR)*.

Li, Y. and Shapiro, A. (2023). Rectangularity and duality of distributionally robust Markov decision processes. *arXiv preprint arXiv:2308.11139*.

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

Lim, S. H., Xu, H., and Mannor, S. (2013). Reinforcement learning in robust Markov decision processes. *Conference on Neural Information Processing Systems (NeurIPS)*.

Liu, Z., Bai, Q., Blanchet, J., Dong, P., Xu, W., Zhou, Z., and Zhou, Z. (2022). Distributionally robust q-learning. *International Conference on Machine Learning (ICML)*.

Liu, Z. and Xu, P. (2024). Distributionally robust off-dynamics reinforcement learning: Provable efficiency with linear function approximation. *arXiv preprint arXiv:2402.15399*.

Ma, X., Liang, Z., Blanchet, J., Liu, M., Xia, L., Zhang, J., Zhao, Q., and Zhou, Z. (2022). Distributionally robust offline reinforcement learning with linear function approximation. *arXiv preprint arXiv:2209.06620*.

Mankowitz, D. J., Levine, N., Jeong, R., Abdolmaleki, A., Springenberg, J. T., Shi, Y., Kay, J., Hester, T., Mann, T., and Riedmiller, M. (2019). Robust reinforcement learning for continuous control with model misspecification. *International Conference on Learning Representations (ICLR)*.

Mannor, S., Mebel, O., and Xu, H. (2016). Robust mdps with k-rectangular uncertainty. *Mathematics of Operations Research (INFORMS)*.

Mehta, V., Paria, B., Schneider, J., Ermon, S., and Neiswanger, W. (2021). An experimental design perspective on model-based reinforcement learning. *International Conference on Learning Representations (ICLR)*.

Nguyen, T., Gupta, S., Ha, H., Rana, S., and Venkatesh, S. (2020). Distributionally robust Bayesian quadrature optimization. *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Nilim, A. and El Ghaoui, L. (2005). Robust control of Markov decision processes with uncertain transition matrices. *Operations Research (INFORMS)*.

Panaganti, K. and Kalathil, D. (2022). Sample complexity of robust reinforcement learning with a generative model. *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Panaganti, K., Xu, Z., Kalathil, D., and Ghavamzadeh, M. (2022). Robust reinforcement learning using offline data. *Conference on Neural Information Processing Systems (NeurIPS)*, 35:32211–32224.

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

Peng, X. B., Andrychowicz, M., Zaremba, W., and Abbeel, P. (2018). Sim-to-real transfer of robotic control with dynamics randomization. *IEEE International Conference on Robotics and Automation (ICRA)*.

Petrik, M. and Russel, R. H. (2019). Beyond confidence regions: Tight Bayesian ambiguity sets for robust MDPs. *Conference on Neural Information Processing Systems (NeurIPS)*.

Pinneri, C., Sawant, S., Blaes, S., Achterhold, J., Stueckler, J., Rolinek, M., and Martius, G. (2020). Sample-efficient cross-entropy method for real-time planning. *arXiv preprint arXiv:2008.06389*.

Pinto, L., Davidson, J., Sukthankar, R., and Gupta, A. (2017). Robust adversarial reinforcement learning. *International Conference on Machine Learning (ICML)*.

Rasmussen, C. E. and Williams, C. (2006). Gaussian processes for machine learning, vol. 1.

Rastogi, D., Koryakovskiy, I., and Kober, J. (2018). Sample-efficient reinforcement learning via difference models. *Machine Learning in Planning and Control of Robot Motion Workshop at ICRA*.

Roy, A., Xu, H., and Pokutta, S. (2017). Reinforcement learning under model mismatch. *Conference on Neural Information Processing Systems (NeurIPS)*.

Shapiro, A. (2017). Distributionally robust stochastic programming. *SIAM Journal on Optimization*.

Shariff, R. and Szepesvári, C. (2020). Efficient planning in large MDPs with weak linear function approximation. *Conference on Neural Information Processing Systems (NeurIPS)*.

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

Shi, L., Li, G., Wei, Y., Chen, Y., Geist, M., and Chi, Y. (2024). The curious price of distributional robustness in reinforcement learning with a generative model. *Conference on Neural Information Processing Systems (NeurIPS)*.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. (2009). Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*.

Tamar, A., Mannor, S., and Xu, H. (2014). Scaling up robust mdps using function approximation. *International Conference on Machine Learning (ICML)*.

Tkachuk, V., Bakhtiari, S. A., Kirschner, J., Jusup, M., Bogunovic, I., and Szepesvári, C. (2023). Efficient planning in combinatorial action spaces with applications to cooperative multi-agent reinforcement learning. *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Vakili, S., Bouziani, N., Jalali, S., Bernacchia, A., and Shiu, D.-s. (2021). Optimal order simple regret for Gaussian process bandits. *Conference on Neural Information Processing Systems (NeurIPS)*.

Wang, S., Si, N., Blanchet, J., and Zhou, Z. (2023a). A finite sample complexity bound for distributionally robust q-learning. *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Wang, S., Si, N., Blanchet, J., and Zhou, Z. (2023b). On the foundation of distributionally robust reinforcement learning. *arXiv preprint arXiv:2311.09018*.

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

Wang, Y. and Zou, S. (2021). Online robust reinforcement learning with model uncertainty. *Conference on Neural Information Processing Systems (NeurIPS)*.

Wiesemann, W., Kuhn, D., and Rustem, B. (2013). Robust Markov decision processes. *Mathematics of Operations Research (INFORMS)*.

Wulfmeier, M., Posner, I., and Abbeel, P. (2017). Mutual alignment transfer learning. *Conference on Robot Learning*.

Xu, H. and Mannor, S. (2010). Distributionally robust Markov decision processes. *Conference on Neural Information Processing Systems (NeurIPS)*.

Xu, Z., Panaganti, K., and Kalathil, D. (2023). Improved sample complexity bounds for distributionally robust reinforcement learning. *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Yang, W., Wang, H., Kozuno, T., Jordan, S. M., and Zhang, Z. (2023). Avoiding model estimation in

robust Markov decision processes with a generative model. *arXiv preprint arXiv:2302.01248*.

Yang, W., Zhang, L., and Zhang, Z. (2022). Toward theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics, Institute of Mathematical Statistics*.

Yin, D., Hao, B., Abbasi-Yadkori, Y., Lazić, N., and Szepesvári, C. (2022). Efficient local planning with linear function approximation. *International Conference on Algorithmic Learning Theory*.

Yu, P. and Xu, H. (2015). Distributionally robust counterpart in Markov decision processes. *IEEE Transactions on Automatic Control*.

Yu, Z., Dai, L., Xu, S., Gao, S., and Ho, C. P. (2024). Fast bellman updates for Wasserstein distributionally robust mdps. *Conference on Neural Information Processing Systems (NeurIPS)*.

Zhang, H., Chen, H., Xiao, C., Li, B., Liu, M., Boning, D., and Hsieh, C.-J. (2020). Robust deep reinforcement learning against adversarial perturbations on state observations. *Conference on Neural Information Processing Systems (NeurIPS)*.

Zhang, S., Hu, Y., Zhang, L., and He, N. (2024). Generalization bounds of nonconvex-(strongly)-concave stochastic minimax optimization. *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Zhou, Z., Zhou, Z., Bai, Q., Qiu, L., Blanchet, J., and Glynn, P. (2021). Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] Included in the supplementary material

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] Included in the supplementary material
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes] Included in the supplementary material
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Theoretical Guarantees of Maximum Variance Reduction (MVR)

We formally introduce the Gaussian process model in Appendix A.1. In Appendix A.2, we describe the confidence bound results from Vakili et al. (2021) and adapt them to the case of multi-output GP models. Finally, in Appendix A.3, we provide sample complexity guarantees for the MVR algorithm.

We recall the introduced notation $\mathcal{X} = \mathcal{S} \times \mathcal{A}$ and remark that we use both (s_i, a_i) and x_i interchangeably in this section.

A.1 Gaussian Process Model

Gaussian process (GP) is a non-parametric model that is often used to express uncertainty over functions on any set (e.g., RKHS). They allow to tractably construct posterior distribution over functions in the set to estimate the unknown non-linear function $\tilde{f} : \mathcal{X} \to \mathbb{R}$ given data containing samples from function $\tilde{f}$. It follows the Bayesian methodology of calculating posterior given the prior and assumes that the function values at any finite subset of the domain $\mathcal{X}$ follow the multivariate Gaussian distribution. One specifies a GP by a prior mean function and a covariance function usually defined using a kernel $k(x, x')$ where $x, x' \in \mathcal{X}$.

Assuming that the samples of $\tilde{f} : \mathcal{X} \to \mathbb{R}$ are noisy measurements of the underlying true function $\tilde{f}$ with i.i.d. Gaussian noise $\mathcal{N}(0, \lambda)$, the posterior mean and covariance function of the posterior distribution can be explicitly calculated. In essence, for $\{x_1, \ldots, x_N\} \in \mathcal{X}$ and $y_n = \tilde{f}(x_n) + \omega_n$, the posterior mean, covariance and variance are given by:

$$\mu_n(x) = k_n(x)(K_n + I_n\lambda)^{-1} y_n, \tag{16}$$

$$k_n(x, x') = k(x, x') - k_n(x)(K_n + I_n\lambda)^{-1} k_n^T(x'),$$

$$\sigma_n^2(x) = k_n(x, x). \tag{17}$$

Here K_n denotes the covariance matrix whose entries are $[K_n]_{i,j} = k(x_i, x_j)$ with $x_i, x_j \in \{x_1, \cdots, x_N\}$ and $k_n(x) = [k(x, x_1), \ldots, k(x, x_N)]$ denotes the covariance vector whose entries are the covariance between x and x_j for all $x_j \in \{x_1, \ldots, x_N\}$. The $n \times n$ identity matrix is denoted as I_n.

We consider multi-output GPs to model the unknown function f that outputs states in $\mathbb{R}^d$. (see Section 3). Similar to Equation (16) and Equation (17), we get analogous expressions for the multi-output case in Equation (5) and Equation (6).

A.2 Non-adaptive Multi-output Confidence Bounds

Our Algorithm 1 uses the maximum variance reduction rule to learn about the transition dynamics. As seen in our analysis (see Theorem 2), we are interested in constructing confidence intervals for f only at the end of n iterations (i.e., after taking n samples), and hence, we do not require anytime confidence bounds (e.g., as in Srinivas et al. (2009)). Moreover, in our algorithm, the current decision (s_i, a_i) does not depend on the previous noise realizations. By focusing on the single-output case first, the following confidence lemma from Vakili et al. (2021), can be used to construct confidence intervals with $\beta(\delta)$ independent of n which holds w.h.p. for a fixed $x \in \mathcal{X}$:

Lemma 3. *Given n noisy observations of $\tilde{f} : \mathcal{X} \to \mathbb{R}$ with $\|\tilde{f}\|_k \leqslant B$ where noise $\{\omega_1, \cdots, \omega_n\}$ is independent of inputs $\{x_1, \cdots x_n\}$, for $\beta(\delta) = B + \frac{\sigma}{\lambda}\sqrt{2\log(2/\delta)}$, and μ_n, σ_n as defined in Equation (16) and Equation (17), the following holds for a fixed $x \in \mathcal{X}$ with probability at least $1 - \delta$,*

$$|\tilde{f}(x) - \mu_n(x)| \leqslant \beta(\delta)\sigma_n(x).$$

To extend this result over the entire input set $x \in \mathcal{X}$, the authors in Vakili et al. (2021) use a discretization assumption which ensures that there exists a discretization $\mathcal{D}_n$ such that $\tilde{f}(x) - \tilde{f}([x]_n) \leqslant \frac{1}{\sqrt{n}}$, where $[x]_n = \arg\min_{x' \in \mathcal{D}_n} \|x - x'\|_2$ and $|\mathcal{D}_n| \leqslant CB^d n^{d/2}$ for C being independent of n and B (RKHS norm bound). Consequently, they obtain the following lemma providing uniform confidence bounds:

Lemma 4. *((Vakili et al., 2021, Theorem-3)) Given n noisy observations of $\tilde{f} : \mathcal{X} \to \mathbb{R}$, $\mathcal{X} \subset \mathbb{R}^d$ satisfying $\|\tilde{f}\|_k \leqslant B$ where noise $\{\omega_1, \cdots, \omega_n\}$ is independent of inputs $\{x_1, \cdots x_n\} \subset \mathcal{X}$ and when there exists discretization*

$\mathcal{D}_n$ of $\mathcal{X}$ with $|\mathcal{D}_n| \leqslant CB^d n^{d/2}$, for $\beta(\delta) = B + \frac{\sigma}{\lambda}\sqrt{2\log(2/\delta)}$ and $\beta_n(\delta) = 2B + \beta(\frac{\delta}{3C(B+\sqrt{n}\beta(2\delta/3n))^d n^{d/2}})$, μ_n, σ_n as defined in Equation (16) *and Equation* (17)*, the following holds for all $x \in \mathcal{D}_n$ with probability at least $1-\delta$,*

$$|\tilde{f}(x) - \mu_n(x)| \leqslant \beta_n(\delta)\sigma_n(x).$$

To extend this result to multiple dimensions as required in our work, we take the same discretization assumption as in Vakili et al. (2021). But considering the multi-output definition of f, we define the modified state-action space $\overline{\mathcal{X}}$. This is in line with Chowdhury and Gopalan (2019), which also has a similar multi-output setting. We define the modified state-action space as $\overline{\mathcal{X}} := \mathcal{S} \times \mathcal{A} \times \{1, 2, \cdots, d\}$ where the last dimension $i \in \{1, 2, \cdots, d\}$ incorporates the index of the output vector, in the sense that $f(\cdot,\cdot) = (\tilde{f}(\cdot,\cdot,1), \cdots, \tilde{f}(\cdot,\cdot,d))$ where $\tilde{f} : \overline{\mathcal{X}} \to \mathbb{R}$. We then detail the discretization assumption as in Vakili et al. (2021) w.r.t. $\tilde{f}$ (see also Section 2 for more details).

Assumption 1. *For every $n \in \mathbb{N}$ and $\tilde{f} \in \mathcal{H}_k(\mathcal{S} \times \mathcal{A} \times \mathcal{I})$ there exists a discretization $\mathcal{D}_n(\mathcal{S} \times \mathcal{A})$ of $\mathcal{S} \times \mathcal{A}$ such that $\tilde{f}(s,a,i) - \tilde{f}([s,a]_n, i) \leqslant \frac{1}{\sqrt{n}}$, where $[s,a]_n = \arg\min_{(s',a') \in \mathcal{D}_n(\mathcal{S}\times\mathcal{A})} \|(s,a) - (s',a')\|_2$, $i \in \mathcal{I}$, and $|\mathcal{D}_n(\mathcal{S} \times \mathcal{A})| \leqslant CB^p n^{p/2}$ ($|\mathcal{D}_n(\mathcal{S} \times \mathcal{A} \times \mathcal{I})| \leqslant CB^p n^{p/2} d$) for C being independent of n and B, and $\mathcal{S} \times \mathcal{A} \subset \mathbb{R}^p$.*

Assumption 1 allows us to provide bounds for $\|f(s,a) - \mu_n(s,a)\|_2$ for all $(s,a) \in \mathcal{S}$ using Lemma 4. Note that Assumption 1 does not discretize the modified state-action space ($\overline{\mathcal{X}} = \mathcal{S} \times \mathcal{A} \times \{1, 2, \cdots, d\}$) but instead discretizes $\mathcal{S} \times \mathcal{A}$ for each $i \in \mathcal{I}$. Hence, $|\mathcal{D}_n(\mathcal{S} \times \mathcal{A} \times \mathcal{I})| \leqslant CB^p n^{p/2} d$, and $\beta_n(\delta)$ will change accordingly. We describe the following lemma detailing the same.

Lemma 5. *Under Assumption 1 with $\beta_n(\delta)$ as in Lemma 4 and training a Gaussian process model on observations up to iteration n ($\{s_1, \cdots, s_n\}$) and their corresponding inputs ($\{(s_0,a_0), \cdots, (s_{n-1}, a_{n-1})\}$), it holds with probability at least $1-\delta$,*

$$\|f(s,a) - \mu_n(s,a)\|_2 \leqslant \beta_n(\delta)\sqrt{d}\|\sigma_n([s,a]_n)\|_2 + \frac{2d}{\sqrt{n}},$$

uniformly for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $[s,a]_n = \arg\min_{(s',a') \in \mathcal{D}_n(\mathcal{S}\times\mathcal{A})} \|(s,a) - (s',a')\|_2$.

Proof. For any $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}
&\|f(s,a) - \mu_n(s,a)\|_2 \\
&= \sqrt{\sum_{i=1}^{d} (\tilde{f}(s,a,i) - \mu_n(s,a,i))^2} && (18)\\
&= \sqrt{\sum_{i=1}^{d} |\tilde{f}(s,a,i) - \tilde{f}([s,a]_n,i) + \tilde{f}([s,a]_n,i) - \mu_n([s,a]_n,i) + \mu_n([s,a]_n,i) - \mu_n(s,a,i)|^2} \\
&\leqslant \sum_{i=1}^{d} \Big(|\tilde{f}(s,a,i) - \tilde{f}([s,a]_n,i)| + |\tilde{f}([s,a]_n,i) - \mu_n([s,a]_n,i)| + |\mu_n([s,a]_n,i) - \mu_n(s,a,i)| \Big) && (19)\\
&\leqslant \left(\sum_{i=1}^{d} (|\tilde{f}([s,a]_n,i) - \mu_n([s,a]_n,i)|) \right) + \frac{2d}{\sqrt{n}} && (20)\\
&\leqslant \beta_n(\delta) \left(\sum_{i=1}^{d} (\sigma_n([s,a]_n,i)) \right) + \frac{2d}{\sqrt{n}} && (21)\\
&\leqslant \beta_n(\delta)\sqrt{d} \sqrt{\sum_{i=1}^{d} (\sigma_n([s,a]_n,i))^2} + \frac{2d}{\sqrt{n}} && (22)\\
&\leqslant \beta_n(\delta)\sqrt{d}\|\sigma_n([s,a]_n)\|_2 + \frac{2d}{\sqrt{n}}. && (23)
\end{aligned}$$

In Equation (19), Equation (22) we use $\|x\|_2 \leqslant \|x\|_1 \leqslant \sqrt{d}\|x\|_2$. And Equation (20) and Equation (21) follow from Assumption 1 (since $\tilde{f}, \mu_n \in \mathcal{H}_k(\mathcal{S} \times \mathcal{A} \times \mathcal{I})$) and Lemma 4, respectively.

□

A.3 Sample Complexity Guarantees

Our objective is to obtain a uniform upper bound on the model precision $\|\mu_n(s,a) - f(s,a)\|_2$ for all state-action pairs (s,a) while accounting for the errors induced by discretization. Here, $\mu_n(\cdot,\cdot)$ is obtained from Algorithm 1. We achieve this by using Lemma 5 to obtain a bound in terms of maximum information gain (Equation (9)).

Lemma 1. *For $\beta_n(\delta)$ set as in Lemma 4 and $\mathcal{I}_d$ denoting $\{1, 2, \cdots, d\}$, the MVR algorithm (Algorithm 1) outputs the dynamics estimate $\hat{f}_n(\cdot,\cdot) = \mu_n(\cdot,\cdot)$ such that the following holds uniformly for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ with probability at least $1-\delta$,*

$$\|\mu_n(s,a) - f(s,a)\|_2 \leqslant \mathcal{O}\Big(\tfrac{\beta_n(\delta)2ed}{\sqrt{n}}\sqrt{\Gamma_{nd}(\mathcal{S}\times\mathcal{A}\times\mathcal{I}_d)}\Big).$$

Proof. From Lemma 5, it holds that with probability at least $1-\delta$ uniformly for all $(s,a) \in \mathcal{S}\times\mathcal{A}$:

$$\begin{aligned}
\|\mu_n(s,a) - f(s,a)\|_2 &\leqslant \beta_n(\delta)\sqrt{d}\|\sigma_n([s,a]_n)\|_2 + \frac{2d}{\sqrt{n}} \\
&\leqslant \beta_n(\delta)\sqrt{d}\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\|\sigma_n(s,a)\|_2 + \frac{2d}{\sqrt{n}} \\
&\leqslant \beta_n(\delta)\sqrt{d}\|\sigma_n(s_n,a_n)\|_2 + \frac{2d}{\sqrt{n}} \\
&\leqslant \frac{2d}{\sqrt{n}} + \frac{\beta_n(\delta)}{n}\sqrt{d}\sum_{j=1}^{n}\|\sigma_j(s_n,a_n)\|_2 \\
&\leqslant \frac{2d}{\sqrt{n}} + \frac{\beta_n(\delta)}{n}\sqrt{d}\sum_{j=1}^{n}\|\sigma_j(s_j,a_j)\|_2 \qquad (24) \\
&\leqslant \frac{\beta_n(\delta)}{\sqrt{n}}\sqrt{d}\sqrt{\sum_{j=1}^{n}\|\sigma_j(s_j,a_j)\|_2^2} + \frac{2d}{\sqrt{n}} \\
&\leqslant \frac{\beta_n(\delta)2ed}{\sqrt{n}}\sqrt{\Gamma_{nd}(\mathcal{S}\times\mathcal{A}\times\mathcal{I}_d)} + \frac{2d}{\sqrt{n}} \qquad (25) \\
&= \mathcal{O}\Big(\frac{\beta_n(\delta)2ed}{\sqrt{n}}\sqrt{\Gamma_{nd}(\mathcal{S}\times\mathcal{A}\times\mathcal{I}_d)}\Big). \qquad (26)
\end{aligned}$$

Here, Equation (24) follows from the decision rule in line-4 of Algorithm 1 and Equation (25) is obtained using standard bound for the sum of variances in the case of multi-output GPs from Curi et al. (2021, Lemma-7) and Chowdhury and Gopalan (2019, Lemma-11). □

B Sample Complexity Bounds for KL Uncertainty Sets

Theorem 2. *(Sample Complexity of MVR under KL uncertainty set) Consider a robust MDP with nominal transition dynamics f satisfying the regularity assumptions from Section 2 and with uncertainty set defined as in Equation (2) w.r.t. KL divergence. For π^* denoting the robust optimal policy w.r.t. nominal transition dynamics f and $\hat{\pi}_N$ denoting the robust optimal policy w.r.t. learned nominal transition dynamics $\hat{f}_N$ via MVR (Algorithm 1), and $\delta \in (0,1)$, $\epsilon \in (0, \frac{1}{1-\gamma})$, it holds that $\max_s |V^R_{\hat{\pi}_N,f}(s) - V^R_{\pi^*,f}(s)| \leqslant \epsilon$ with probability at least $1-\delta$ for any N such that*

$$N = \mathcal{O}\Big(e^{\frac{2-\gamma}{(1-\gamma)\alpha_{\mathrm{kl}}}} \frac{\gamma^2 \beta_N^2(\delta) d^2 \Gamma_{Nd}}{(1-\gamma)^4 \rho^2 \epsilon^2}\Big). \tag{27}$$

Proof. **Step (i):** As detailed in the proof outline of Section 4, in order to bound $|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)|$, we begin by adding and subtracting $V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s)$ which is the robust value function w.r.t. the nominal transition dynamics $\hat{f}_n$ and its corresponding optimal policy $\hat{\pi}_n$. Then, we split the difference into two terms as follows:

$$|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)| = \underbrace{|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s)|}_{(i)} + \underbrace{|V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)|}_{(ii)}. \tag{28}$$

In order to not disturb the flow of the proof we bound (i) and (ii) separately Lemma 6 and Lemma 7 respectively. From Lemma 6, we obtain that

$$\begin{aligned}(i) &\leqslant \max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s)\Big| \\ &\leqslant \frac{\gamma}{1-\gamma} \max_s \Big| \inf_{\mathrm{KL}(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big] - \inf_{\mathrm{KL}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big]\Big|. \end{aligned} \tag{29}$$

And from Lemma 7, we obtain that

$$\begin{aligned}(ii) &\leqslant \max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)\Big| \\ &\leqslant \frac{\gamma}{1-\gamma} \max_s \Big| \inf_{\mathrm{KL}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\pi^*,f}(s')\Big] - \inf_{\mathrm{KL}(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\pi^*,f}(s')\Big]\Big|. \end{aligned} \tag{30}$$

Note that both these terms in Equations (29) and (30) are of the form mentioned in the **Step (i)** of Section 4.

Step (ii): Next, corresponding to **Step (ii)** of the proof outline in Section 4, we use Lemma 2 to bound Equations (29) and (30). Denote $M := \frac{1}{1-\gamma} \geqslant \max_s V^{\mathrm{R}}_{\pi}(s)$ for convenience. Using Equation (29) and Lemma 9 (internally using Lemma 2), conditioned on the event of Lemma 9 holding true, it holds that

$$\begin{aligned}(i) &\leqslant \max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s)\Big| \\ &\leqslant \frac{1}{1-\gamma} \max_s \Big|\gamma \inf_{\mathrm{KL}(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big] - \gamma \inf_{\mathrm{KL}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big]\Big| \\ &\leqslant \max_{s,a}\left(2\gamma\frac{M^2}{\rho}e^{\frac{M}{\alpha}} \max_{\alpha\in[\underline{\alpha},\frac{M}{\rho}]} \Big|\mathbb{E}_{s'\sim P_{\hat{f}_n}(s,a)}[e^{\frac{-V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')}{\alpha}}] - \mathbb{E}_{s'\sim P_f(s,a)}[e^{\frac{-V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')}{\alpha}}]\Big|\right). \end{aligned} \tag{31}$$

$$(i) \leqslant \max_{V(\cdot)\in\mathcal{V}} \max_{s,a}\left(2\gamma\frac{M^2}{\rho}e^{\frac{M}{\alpha}} \max_{\alpha\in[\underline{\alpha},\frac{M}{\rho}]} \Big|\mathbb{E}_{s'\sim P_{\hat{f}_n}(s,a)}[e^{\frac{-V(s')}{\alpha}}] - \mathbb{E}_{s'\sim P_f(s,a)}[e^{\frac{-V(s')}{\alpha}}]\Big|\right). \tag{32}$$

We can bound (ii) similarly.

$$(ii) \leqslant \max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)\Big| \tag{33}$$

$$\leqslant \max_{V(\cdot)\in\mathcal{V}} \max_{s,a}\left(2\gamma\frac{M^2}{\rho}e^{\frac{M}{\alpha}} \max_{\alpha\in[\underline{\alpha},\frac{M}{\rho}]} \Big|\mathbb{E}_{s'\sim P_{\hat{f}_n}(s,a)}[e^{\frac{-V(s')}{\alpha}}] - \mathbb{E}_{s'\sim P_f(s,a)}[e^{\frac{-V(s')}{\alpha}}]\Big|\right). \tag{34}$$

Step (iii): Next, we want to utilize the learning error bound (Equation (26)) that bounds the difference between the means of true nominal transition dynamics P_f and learned nominal transition dynamics $P_{\hat{f}_n}$ to bound Equations (32) and (34).

We begin by bounding the difference $\left|\mathbb{E}_{s'\sim P_{\hat{f}_n}(s,a)}[e^{\frac{-V(s')}{\alpha}}] - \mathbb{E}_{s'\sim P_f(s,a)}[e^{\frac{-V(s')}{\alpha}}]\right|$, by the difference in means of P_f and $P_{\hat{f}_n}$ in Lemma 10. Since Equation (32) has a max over all value functions, we introduce a covering number argument in Lemma 12 to reform it to a max over the functions in the $\zeta-$covering set. We then use Lemma 10 to obtain bounds in terms of maximum information gain Γ_{Nd} (Equation (9)) and ζ. Further details regarding the covering number argument are deferred to Lemma 12. Then, we apply the result of Lemma 12 with $\zeta = 1$ (defined in Lemma 12) on Equation (32). Then, it holds that

$$(i) \leqslant \max_s \left|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s)\right| = \mathcal{O}\left(2\tfrac{M^2}{\rho} e^{\frac{M}{\alpha_{kl}}} e^{\frac{1}{\alpha_{kl}}} \frac{\beta_n(\delta)\sqrt{2ed^2\Gamma_{nd}}}{\sigma\sqrt{n}}\right), \tag{35}$$

where α_{kl} is a problem-dependent constant denoting the minimum value of $\underline{\alpha}$ defined in Lemma 9. A similar constant also appears in the sample complexity bounds provided in Panaganti and Kalathil (2022); Zhou et al. (2021). Note that β_n, which appears in Lemma 3, has a logarithmic dependence on n. Similarly, from Equation (34) and Lemmas 10 and 12, we obtain

$$(ii) \leqslant \max_s \left|V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)\right| = \mathcal{O}\Big(2\gamma\tfrac{M^2}{\rho} e^{\frac{M}{\alpha_{kl}}} e^{\frac{1}{\alpha_{kl}}} \frac{\beta_n(\delta)\sqrt{2ed^2\Gamma_{nd}}}{\sigma\sqrt{n}}\Big). \tag{36}$$

Note that we want to bound $V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\pi^*,f}(s) = (i) + (ii)$ over all $s \in \mathcal{S}$. Using $\max_s \left|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)\right| \leqslant \max_s \left|V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)\right| + \max_s \left|V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)\right|$ and substituting M by $1/(1-\gamma)$, we obtain from Equation (35) and Equation (36)

$$\max_s \left|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)\right| = \mathcal{O}\Big(\gamma e^{\frac{1}{(1-\gamma)\alpha_{kl}}} e^{\frac{1}{\alpha_{kl}}} \frac{\beta_n(\delta) d\sqrt{2e\Gamma_{nd}}}{(1-\gamma)^2\rho\sigma\sqrt{n}}\Big).$$

Finally, to ensure that $\max_s |V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)| \leqslant \epsilon$, it suffices to have

$$\max_s \left|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)\right| = \mathcal{O}\Big(\gamma e^{\frac{1}{(1-\gamma)\alpha_{kl}}} e^{\frac{1}{\alpha_{kl}}} \frac{\beta_n(\delta) d\sqrt{2e\Gamma_{nd}}}{(1-\gamma)^2\rho\sigma\sqrt{n}}\Big) = \epsilon.$$

By inverting the previously obtained result, we arrive at

$$n = \mathcal{O}\Big(e^{\frac{2}{(1-\gamma)\alpha_{kl}}} e^{\frac{2}{\alpha_{kl}}} \frac{\gamma^2\beta_n^2(\delta) d^2\Gamma_{nd}}{(1-\gamma)^4\rho^2\epsilon^2}\Big).$$

□

Lemma 6. *(Simplification using robust Bellman equation) Denote* $(i) := \left|V^R_{\hat{\pi}_n,f}(s) - V^R_{\hat{\pi}_n,\hat{f}_n}(s)\right|$ *for* $V^R_{\hat{\pi}_n,f}$ *being the robust value function of policy* $\hat{\pi}_n$ *w.r.t. true nominal transition dynamics* f *and* $V^R_{\hat{\pi}_n,\hat{f}_n}$ *being the robust value function of policy* $\hat{\pi}_n$ *w.r.t. learned nominal transition dynamics* f*. Then the following holds,*

$$\begin{aligned}(i) &= \left|V^R_{\hat{\pi}_n,f}(s) - V^R_{\hat{\pi}_n,\hat{f}_n}(s)\right| \\ &\leqslant \max_s \left|V^R_{\hat{\pi}_n,f}(s) - V^R_{\hat{\pi}_n,\hat{f}_n}(s)\right| \\ &\leqslant \frac{\gamma}{1-\gamma} \max_s \Big| \inf_{\mathrm{D}(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^R_{\hat{\pi}_n,f}(s')\Big] - \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^R_{\hat{\pi}_n,f}(s')\Big]\Big|. \end{aligned} \tag{37}$$

Proof. Since both the quantities are w.r.t. the same policy, using the definition of the robust Q-function and the robust Bellman equation (see Equation (4)), we obtain:

$$(i) = |V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s)| \tag{38}$$

$$
\begin{aligned}
&= |Q^{\mathrm{R}}_{\hat{\pi}_n,f}(s,\hat{\pi}_n(s)) - Q^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s,\hat{\pi}_n(s))| \\
&= |r(s,\hat{\pi}_n(s)) - r(s,\hat{\pi}_n(s)) \\
&\quad + \gamma \inf_{\mathrm{D}(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big] - \gamma \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s')\Big]| \\
&= |\gamma \inf_{\mathrm{D}(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big] - \gamma \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s')\Big]| \qquad (39)
\end{aligned}
$$

Adding and subtracting $\gamma \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big]$ to Equation (39), we obtain the following two terms:

$$
(i_a) = |\gamma \inf_{\mathrm{D}(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big] - \gamma \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big]|,
$$

$$
(i_b) = |\gamma \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big] - \gamma \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s')\Big]|.
$$

Now, we use Lemma 8 to bound (i_b). We have:

$$
\begin{aligned}
(i_b) &= |\gamma \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big] - \gamma \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s')\Big]| \\
&\overset{Lemma\ 8}{\leqslant} \gamma \max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s)\Big| \qquad (Lemma\ 8). \qquad (40)
\end{aligned}
$$

Plugging Equation (40) into Equation (38) and using the fact that $(i) = (i_a) + (i_b)$, we have

$$
\begin{aligned}
(i) &= |V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s)| \qquad (41) \\
&\leqslant (i_a) + \gamma \max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s)\Big| \\
&= |\gamma \inf_{\mathrm{D}(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big] - \gamma \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big]| \\
&\quad + \gamma \max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s)\Big|. \qquad (42)
\end{aligned}
$$

Taking maximum over states in Equation (41) and Equation (42) we have

$$
\begin{aligned}
&\max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s)\Big| \\
&\leqslant \max_s \Big|\gamma \inf_{\mathrm{D}(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big] - \gamma \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big]\Big| \\
&\quad + \gamma \max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s)\Big|.
\end{aligned}
$$

Moving $\gamma \max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s)\Big|$ to the LHS and dividing $(1-\gamma)$ on both sides, it holds that

$$
\begin{aligned}
(i) &\leqslant \max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s)\Big| \\
&\leqslant \frac{\gamma}{1-\gamma} \max_s \Big|\inf_{\mathrm{D}(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big] - \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big]\Big|. \qquad (43)
\end{aligned}
$$

□

Lemma 7. *(Simplification using robust Bellman equation) Denote* $(ii) := \Big|V^{R}_{\hat{\pi}_n,\hat{f}_n}(s) - V^{R}_{\pi^*,f}(s)\Big|$ *for* $V^{R}_{\hat{\pi}_n,\hat{f}_n}$ *being the robust value function of policy* $\hat{\pi}_n$ *w.r.t. learned nominal transition dynamics* $\hat{f}_n$ *and* $V^{R}_{\pi^*,f}$ *being the robust value function of policy* π^* *w.r.t. true nominal transition dynamics* f*. Then the following holds,*

$$
(ii) = \Big|V^{R}_{\hat{\pi}_n,\hat{f}_n}(s) - V^{R}_{\pi^*,f}(s)\Big|
$$

$$\leqslant \max_s \Big| V^R_{\hat{\pi}_n, \hat{f}_n}(s) - V^R_{\pi^*, f}(s) \Big|$$
$$\leqslant \frac{\gamma}{1-\gamma} \max_s \Big| \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^R_{\pi^*,f}(s')\Big] - \inf_{\mathrm{D}(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^R_{\pi^*,f}(s')\Big]\Big|. \tag{44}$$

Proof. We first note that $Q^{\mathrm{R}}_{\pi^*,f}(s,\hat{\pi}_n(s)) \leqslant Q^{\mathrm{R}}_{\pi^*,f}(s,\pi^*(s))$ as π^* is the robust optimal policy for the nominal transition dynamics f (see Equation (3)). As a result, we have

$$(ii) = |V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)| \tag{45}$$
$$= |Q^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s,\hat{\pi}_n(s)) - Q^{\mathrm{R}}_{\pi^*,f}(s,\pi^*(s))|$$
$$\leqslant |Q^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s,\hat{\pi}_n(s)) - Q^{\mathrm{R}}_{\pi^*,f}(s,\hat{\pi}_n(s))|$$
$$= |r(s,\hat{\pi}_n(s)) - r(s,\hat{\pi}_n(s))$$
$$+ \gamma \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s')\Big] - \gamma \inf_{\mathrm{D}(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\pi^*,f}(s')\Big]|.$$
$$= |\gamma \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s')\Big] - \gamma \inf_{\mathrm{D}(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\pi^*,f}(s')\Big]| \tag{46}$$

Adding and subtracting $\gamma \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\pi^*,f}(s')\Big]$ to Equation (46), we obtain the following two terms:

$$(ii_a) = |\gamma \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s')\Big] - \gamma \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\pi^*,f}(s')\Big]|,$$

$$(ii_b) = |\gamma \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\pi^*,f}(s')\Big] - \gamma \inf_{\mathrm{D}(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\pi^*,f}(s')\Big]|.$$

Now, we use Lemma 8 to bound (ii_a) . We have:

$$(ii_a) = |\gamma \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s')\Big] - \gamma \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\pi^*,f}(s')\Big]|$$
$$\leqslant \gamma \max_s \Big| V^{\mathrm{R}}_{\pi^*,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) \Big|. \tag{47}$$

Plugging Equation (47) into Equation (45) and using the fact that $(ii) = (ii_a) + (ii_b)$, we have

$$(ii) = |V^{\mathrm{R}}_{\pi^*,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s)| \tag{48}$$
$$\leqslant (ii_b) + \max_s \Big| V^{\mathrm{R}}_{\pi^*,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) \Big|$$
$$= |\gamma \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\pi^*,f}(s')\Big] - \gamma \inf_{\mathrm{D}(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\pi^*,f}(s')\Big]| \tag{49}$$
$$+ \gamma \max_s \Big| V^{\mathrm{R}}_{\pi^*,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) \Big|.$$

Taking maximum over states in Equation (48) and Equation (49) and following similar steps as in Equation (43), we have

$$(ii) \leqslant \max_s \Big| V^{\mathrm{R}}_{\pi^*,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) \Big|$$
$$\leqslant \max_s \Big| \gamma \inf_{\mathrm{D}(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big] - \gamma \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big] \Big|$$
$$+ \gamma \max_s \Big| V^{\mathrm{R}}_{\pi^*,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) \Big|$$
$$\leqslant \frac{\gamma}{1-\gamma} \max_s \Big| \inf_{\mathrm{D}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\pi^*,f}(s')\Big] - \inf_{\mathrm{D}(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\pi^*,f}(s')\Big]\Big|. \tag{50}$$

□

Lemma 8. *(from Panaganti and Kalathil (2022, Lemma 1)) Let V_1 and V_2 be two value functions from $\mathcal{S} \to [0, 1/(1-\gamma)]$. Let D be any distance measure between probability distributions (e.g., KL-divergence, χ^2- divergence, or variation distance defined in Equation (2)). The following inequality (1-Lipschitz w.r.t. V) holds true*

$$\Big| \inf_{D(p||P_{\tilde{f}}(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V_1(s')\Big] \quad - \quad \inf_{D(p||P_{\tilde{f}}(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V_2(s')\Big]\Big| \quad \leqslant \quad \max_{s'} |V_2(s') \quad - \quad V_1(s')|.$$

Proof. We want to bound

$$\Big| \inf_{D(p||P_{\tilde{f}}(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V_1(s')\Big] - \inf_{D(p||P_{\tilde{f}}(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V_2(s')\Big]\Big|.$$

Notice that

$$\begin{aligned}
&\inf_{D(p||P_{\tilde{f}}(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V_1(s')\Big] - \inf_{D(p||P_{\tilde{f}}(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V_2(s')\Big] \\
&= \inf_{D(p||P_{\tilde{f}}(s,a))\leqslant\rho} \sup_{D(p'|P_{\tilde{f}}(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V_1(s')\Big] - \mathbb{E}_{s'\sim p'}\Big[V_2(s')\Big] \\
&\geqslant \inf_{D(p||P_{\tilde{f}}(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V_1(s')\Big] - \mathbb{E}_{s'\sim p}\Big[V_2(s')\Big] \\
&= \inf_{D(p||P_{\tilde{f}}(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V_1(s') - V_2(s')\Big],
\end{aligned}$$

where the inequality follows from the property of supremum. By the definition of inf, for any $\epsilon > 0$, there exists some distribution q s.t. $D(q|P_{\tilde{f}}(s,a)) \leqslant \rho$ satisfying

$$\mathbb{E}_{s'\sim q}\Big[V_1(s') - V_2(s')\Big] - \epsilon \leqslant \inf_{D(p||P_{\tilde{f}}(s,a)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V_1(s') - V_2(s')\Big].$$

Then, we have

$$\begin{aligned}
&\inf_{D(p||P_{\tilde{f}}(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V_2(s')\Big] - \inf_{D(p||P_{\tilde{f}}(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V_1(s')\Big] \\
&\leqslant - \inf_{D(p||P_{\tilde{f}}(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V_1(s') - V_2(s')\Big] \\
&\leqslant -\mathbb{E}_{s'\sim q}\Big[V_1(s') - V_2(s')\Big] + \epsilon \\
&\leqslant \mathbb{E}_{s'\sim q}\Big[V_2(s') - V_1(s')\Big] + \epsilon \\
&\leqslant \max_{s'} |V_2(s') - V_1(s')| + \epsilon. \qquad (51)
\end{aligned}$$

Let $\epsilon \to 0$, we obtain one side of the desired bound.

One can similarly bound $\inf_{D(p||P_{\tilde{f}}(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V_1(s')\Big] - \inf_{D(p||P_{\tilde{f}}(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V_2(s')\Big]$ by just interchanging V_1 and V_2 everywhere. Combining this argument with Equation (51), we obtain

$$\Big| \inf_{D(p||P_{\tilde{f}}(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V_1(s')\Big] \quad - \quad \inf_{D(p||P_{\tilde{f}}(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V_2(s')\Big]\Big| \quad \leqslant \quad \max_{s'} |V_2(s') \quad - \quad V_1(s')|.$$

□

Lemma 9. *(Simplification using Lemma 2 reformulation) For any value function $V(\cdot) : \mathcal{S} \to [0, 1/(1-\gamma)]$, define the event $\boldsymbol{E}$ as follows:*

$$\max_s \left| \inf_{KL(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V(s')\Big] - \inf_{KL(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V(s')\Big] \right| \leqslant$$
$$\max_{s,a} 2\frac{M}{\rho} e^{\frac{M}{\alpha}} \max_{\alpha\in[\underline{\alpha},\overline{\alpha}]} \left| \mathbb{E}_{s'\sim P_{\hat{f}_{n(s,a)}}}[e^{\frac{-V(s')}{\alpha}}] - \mathbb{E}_{s'\sim P_f(s,a)}[e^{\frac{-V(s')}{\alpha}}] \right|.$$

Then, for any $n > \{\max_{s,a} N'(\rho, P_f(s,a)), \max_{s,a} N''(\rho, P_f(s,a))\}$ *where* $N'(\rho, P_f(s,a)) = \mathcal{O}\Big(\frac{\beta_n^2(\delta) 2ed^2\Gamma_{nd}}{(\frac{\kappa - e^{-\rho}}{2})^2}\Big)$ *and* $N''(\rho, P_f(s,a)) = \mathcal{O}\Big(\frac{4M^2 e^{\frac{2M}{\underline{\alpha}}} \beta_n^2(\delta) 2ed^2\Gamma_{nd}}{(\rho\tau)^2}\Big)$ *with* $\overline{\alpha} = \frac{M}{\rho}$, $M = \frac{1}{1-\gamma}$, κ *defined in Equation* (67), τ *defined in Equation* (70), *and* $\underline{\alpha} = \alpha^*/2$ *defined in Equation* (56), *the event* $\boldsymbol{E}$ *holds true with probability at least* $1-\delta$.

Proof. (A similar proof as in Zhou et al. (2021, Lemma-4)). First note that,

$$\max_s \left| \inf_{KL(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V(s')\Big] - \inf_{KL(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V(s')\Big] \right| \leqslant \max_{s,a} \left| \inf_{KL(p||P_{\hat{f}_n}(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V(s')\Big] - \inf_{KL(p||P_f(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V(s')\Big] \right|. \quad (52)$$

Recall (Hu and Hong, 2013, Theorem-1) for distributionally robust optimization with a random variable X and a random function H. One can rewrite an infinite-dimensional optimization problem as a scalar optimization problem:

$$\sup_{P:KL(p||P_0)\leqslant\rho} \mathbb{E}_{X\sim P}[H(X)] = \inf_{\alpha\geqslant 0}\{\alpha \log(\mathbb{E}_{X\sim P_0}[e^{\frac{H(X)}{\alpha}}]) + \alpha\rho\}. \quad (53)$$

For now, we focus on bounding $\left| \inf_{KL(p||P_{\hat{f}_n}(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V(s')\Big] - \inf_{KL(p||P_f(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V(s')\Big] \right|$ for one particular (s,a). For brevity, we write $P_f(s,a)$ and $P_{\hat{f}_n}(s,a)$ as P_f and $P_{\hat{f}_n}$, respectively. By Equation (53), we have

$$\inf_{P:KL(p||P_f)\leqslant\rho} \mathbb{E}_{s'\sim P}[V(s')] = \max_{\alpha\geqslant 0}\{-\alpha \log(\mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho\}, \quad (54)$$

$$\inf_{P:KL(p||\hat{P}_{\hat{f}_n})\leqslant\rho} \mathbb{E}_{s'\sim P}[V(s')] = \max_{\alpha\geqslant 0}\{-\alpha \log(\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho\}. \quad (55)$$

For the finite state-action space setting, Zhou et al. (2021, Lemma-4) characterizes the property of the optimal α^*. Following a similar proof strategy, we denote

$$\alpha^* = \arg\max_{\alpha\geqslant 0}\{-\alpha \log(\mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho\}, \quad (56)$$

and

$$\hat{\alpha}_n^* = \arg\max_{\alpha\geqslant 0}\{-\alpha \log(\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho\}. \quad (57)$$

To ensure that $\max_{\alpha\geqslant 0}\{-\alpha \log(\mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho\} - \max_{\alpha\geqslant 0}\{-\alpha \log(\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho\}$ is small enough, we need to show that α^* and $\hat{\alpha}_n^*$ are close enough. For this, one considers two different cases, $\alpha^* = 0$ and $\alpha^* > 0$.

Case-1: In Case-1, we investigate the conditions for $\hat{\alpha}_n^* = 0$ given that $\alpha^* = 0$. According to (Hu and Hong, 2013, Proposition-2), for $\alpha^* = 0$ to occur, the random variable $Y := V(s')$ where $s' \sim \mathcal{N}(f(s,a), \sigma^2 I)$ must satisfy three conditions namely, (i) Y must be bounded, (ii) Y must have finite mass at its essential infimum, and (iii) the finite mass at essential infimum should be greater than $e^{-\rho}$. So we want to verify whether these conditions hold true for $\hat{Y}_n := V(s')$ where $s' \sim \mathcal{N}(\hat{f}_n(s,a), \sigma^2 I)$ when Y satisfies these conditions.

We restate definition of the essential infimum for a real-valued random variable Y, denoted as $\mathrm{ESI}(Y)$.

$$\mathrm{ESI}(Y) = \sup\{t \in \mathbb{R} : \mathbb{P}\{Y < t\} = 0\}. \quad (58)$$

We first show that $Y = V(s')$ where $s' \sim \mathcal{N}(f(s,a), \sigma^2 I)$ and $\hat{Y}_n = V(s')$ where $s' \sim \mathcal{N}(\hat{f}_n(s,a), \sigma^2 I)$ have the same essential infimum. By the definition of $\mathrm{ESI}(Y)$, for any $\epsilon > 0$, it holds that

$$\mathbb{P}\{\mathrm{ESI}(Y) \leqslant Y < \mathrm{ESI}(Y) + \epsilon\} > 0, \quad \mathbb{P}\{Y < \mathrm{ESI}(Y)\} = 0. \quad (59)$$

It implies for $Y = V(s')$ and $s' \sim \mathcal{N}(f(s,a), \sigma^2 I)$ that

$$\mathbb{P}_{s' \sim \mathcal{N}(f(s,a),\sigma^2 I)}\{s' \in \mathbb{R}^d : \text{ESI}(Y) \leqslant Y = V(s') < \text{ESI}(Y) + \epsilon\} > 0, \tag{60}$$

$$\mathbb{P}_{s' \sim \mathcal{N}(f(s,a),\sigma^2 I)}\{s' \in \mathbb{R}^d : Y = V(s') < \text{ESI}(Y)\} = 0. \tag{61}$$

It further implies that, the set $\{s' \in \mathbb{R}^d : \text{ESI}(Y) \leqslant V(s') < \text{ESI}(Y) + \epsilon\}$ must have a Lebesgue measure greater than 0 and $\{s' \in \mathbb{R}^d : V(s') < \text{ESI}(Y)\}$ must have a Lebesgue measure equal to 0 since $s' \sim \mathcal{N}(f(s,a), \sigma^2 I)$ is a continuous distribution.

Due to this fact that the set $\{s' \in \mathbb{R}^d : \text{ESI}(Y) \leqslant V(s') < \text{ESI}(Y) + \epsilon\}$ has a Lebesgue measure greater than zero and noting that $\mathcal{N}(\hat{f}_n(s,a), \sigma^2 I)$ is also a continuous distribution with the same support as of $\mathcal{N}(f(s,a), \sigma^2 I)$ (i.e., the probability density function of $\mathcal{N}(\hat{f}_n(s,a), \sigma^2 I)$ is positive whenever probability density function of $\mathcal{N}(f(s,a), \sigma^2 I)$ is positive), it holds that

$$\mathbb{P}_{s' \sim \mathcal{N}(\hat{f}_n(s,a),\sigma^2 I)}\{s' \in \mathbb{R}^d : \text{ESI}(Y) \leqslant \hat{Y}_n = V(s') < \text{ESI}(Y) + \epsilon\} > 0. \tag{62}$$

A similar argument follows for

$$\mathbb{P}_{s' \sim \mathcal{N}(\hat{f}_n(s,a),\sigma^2 I)}\{s' \in \mathbb{R}^d : \hat{Y}_n = V(s') < \text{ESI}(Y)\} = 0. \tag{63}$$

In essence, Equations (62) and (63) imply,

$$\mathbb{P}\{\text{ESI}(Y) \leqslant \hat{Y}_n < \text{ESI}(Y) + \epsilon\} = 0, \quad \mathbb{P}\{\hat{Y}_n < \text{ESI}(Y)\} > 0.$$

Hence, from the definition of $\text{ESI}(\cdot)$ in Equations (58) and (59), we have $\text{ESI}(Y) = \text{ESI}(\hat{Y}_n)$.

As a result, for $\alpha^* = 0$ to occur and for $Y = V(s')(s' \sim \mathcal{N}(f(s,a), \sigma^2 I))$ to have finite mass at the essential infimum (condition-(ii)), i.e, $\mathbb{P}\{Y = \text{ESI}(Y)\} > 0$, it requires that

$$\mathbb{P}_{s' \sim \mathcal{N}(f(s,a),\sigma^2 I)}\{s' \in \mathbb{R}^d : Y = V(s') = \text{ESI}(Y)\} > 0.$$

This will further require that the set $\{s' \in \mathbb{R}^d : Y = V(s') = \text{ESI}(Y)\}$ must have a Lebesgue measure greater than 0. Following a similar argument as to have obtained Equation (62) (the probability density function of $\mathcal{N}(\hat{f}_n(s,a), \sigma^2 I)$ is positive whenever probability density function of $\mathcal{N}(f(s,a), \sigma^2 I)$ is positive), the set $\{s' \in \mathbb{R}^d : Y = V(s') = \text{ESI}(Y)\}$ having Lebesgue measure greater than 0, will imply

$$\mathbb{P}_{s' \sim \mathcal{N}(\hat{f}_n(s,a),\sigma^2 I)}\{s' \in \mathbb{R}^d : \hat{Y}_n = V(s') = \text{ESI}(Y)\} > 0, \tag{64}$$

and

$$\mathbb{P}\{\hat{Y}_n = \text{ESI}(Y)\} > 0 \tag{65}$$

Since $\text{ESI}(Y) = \text{ESI}(\hat{Y}_n)$, Equations (64) and (65) imply

$$\mathbb{P}\{\hat{Y}_n = \text{ESI}(\hat{Y}_n)\} > 0, \tag{66}$$

Hence, if $\mathbb{P}\{Y = \text{ESI}(Y)\} > 0$ holds true, it also holds that $\mathbb{P}\{\hat{Y}_n = \text{ESI}(\hat{Y}_n)\} > 0$. This implies that whenever Y has a finite mass at its essential infimum, $\hat{Y}_n$ also has finite mass at its essential infimum (condition-(ii) satisfied).

But, recall that according to (Hu and Hong, 2013, Proposition-2) for $\alpha^* = 0$ to occur, the finite mass which Y has at its essential infimum should also be greater than $e^{-\rho}$ (condition-(iii)). Hence, one has to check if Y satisfies

$$\mathbb{P}_{s' \sim \mathcal{N}(f(s,a),\sigma^2 I)}\{s' \in \mathbb{R}^d : Y = V(s') = \text{ESI}(Y)\} > e^{-\rho}, \tag{67}$$

what is the condition that Y_n satisfies

$$\mathbb{P}_{s' \sim \mathcal{N}(\hat{f}_n(s,a),\sigma^2 I)}\{s' \in \mathbb{R}^d : \hat{Y}_n = V(s') = \text{ESI}(\hat{Y}_n)\} > e^{-\rho},$$

so that $\hat{\alpha}_n^* = 0$ whenever $\alpha^* = 0$. Denote $\kappa := \mathbb{P}_{s' \sim \mathcal{N}(f(s,a),\sigma^2 I)}\{s' \in \mathbb{R}^d : Y = V(s') = \text{ESI}(Y)\}$, $\kappa_n := \mathbb{P}_{s' \sim \mathcal{N}(\hat{f}_n(s,a),\sigma^2 I)}\{s' \in \mathbb{R}^d : \hat{Y}_n = V(s') = \text{ESI}(\hat{Y}_n)\}$, and $S_{min} := \{s' \in \mathbb{R}^d : V(s') = \text{ESI}(Y) = \text{ESI}(\hat{Y}_n)\}$. If $\kappa > e^{-\rho}$ and $\kappa - \kappa_n \leqslant \frac{\kappa - e^{-\rho}}{2}$, then it will hold that $\kappa_n > e^{-\rho}$.

$$|\kappa - \kappa_n| = \left| \int_{S_{min}} \frac{1}{\sqrt{(2\pi\sigma^2)^d}} (e^{-\frac{\|s' - f(s,a)\|^2}{\sigma^2}} - e^{-\frac{\|s' - \hat{f}_n(s,a)\|^2}{\sigma^2}}) dx \right|$$

$$
\begin{aligned}
&\leqslant \int_{S_{min}} \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \left| e^{-\frac{\|s'-f(s,a)\|^2}{\sigma^2}} - e^{-\frac{\|s'-\hat{f}_n(s,a)\|^2}{\sigma^2}} \right| dx \\
&\leqslant \int_{\mathbb{R}^d} \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \left| e^{-\frac{\|s'-f(s,a)\|^2}{\sigma^2}} - e^{-\frac{\|s'-\hat{f}_n(s,a)\|^2}{\sigma^2}} \right| dx \\
&\leqslant \|f(s,a) - \hat{f}_n(s,a)\|_2 \quad (\textit{Lemma } 10) \\
&\leqslant \mathcal{O}\Big(\frac{\beta_n(\delta)\sqrt{2ed^2\Gamma_{nd}}}{\sqrt{n}}\Big),
\end{aligned}
$$

We need $\mathcal{O}\Big(\frac{\beta_n(\delta)\sqrt{2ed^2\Gamma_{nd}}}{\sqrt{n}}\Big) \leqslant \frac{\kappa - e^{-\rho}}{2}$, which in turn requires $n = \mathcal{O}\Big(\frac{\beta_n^2(\delta)2ed^2\Gamma_{nd}}{(\frac{\kappa-e^{-\rho}}{2})^2}\Big) = N'(\rho, P_f(s,a))$. Hence, for $n > \max_{s,a} N'(\rho, P_f(s,a))$ with probability at least $1-\delta$, it holds that

$$\kappa_n > e^{-\rho},$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ whenever $\kappa > e^{-\rho}$, implying $\alpha_n^* = 0$ whenever $\alpha^* = 0$.

Case-2: Consider the case of $\alpha^* > 0$. The idea is to bound both α^* and $\hat{\alpha}_n^*$ by a set $[\underline{\alpha}, \overline{\alpha}]$ and bound $\max_{\alpha \geqslant 0}\{(-\alpha \log(\mathbb{E}_{s' \sim P_f(s,\pi(s))}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho) - (-\alpha \log(\mathbb{E}_{s' \sim P_{f(s,\pi'(s))}}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho)\}$ for α taking values within set $[\underline{\alpha}, \overline{\alpha}]$. We first provide a upper bound for α^* as $\frac{M}{\rho}$ where $M = \frac{1}{1-\gamma}$ denoting the maximum value of $V(s')$.

$$
\begin{aligned}
\max_{\alpha \geqslant 0}\{-\alpha \log(\mathbb{E}_{s' \sim P_f}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho\} &\geqslant \lim_{\alpha \to 0}[-\alpha \log(\mathbb{E}_{s' \sim P_f}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho] \\
&= \mathrm{ESI}(V(s')|_{s' \sim P_f}) \quad (\textit{Lemma } 11) \\
&\geqslant 0. \qquad (68)
\end{aligned}
$$

Since $\max_s V(s) \leqslant M$, we have

$$-\alpha \log(\mathbb{E}_{s' \sim P_f}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho \leqslant -\alpha \log(e^{\frac{-M}{\alpha}}) - \alpha\rho = M - \alpha\rho.$$

It implies for $\alpha > \frac{M}{\rho}$ that

$$-\alpha \log(\mathbb{E}_{s' \sim P_f}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho < 0. \qquad (69)$$

By Equation (68), since $\max_{\alpha \geqslant 0}\{-\alpha \log(\mathbb{E}_{s' \sim P_f}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho\} \geqslant 0$, $\arg\max_{\alpha \geqslant 0}\{-\alpha \log(\mathbb{E}_{s' \sim P_f}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho\}$ cannot be greater than $\frac{M}{\rho}$ due to Equation (69) holding for all $\alpha > \frac{M}{\rho}$. Hence, we have $\alpha^* \leqslant \frac{M}{\rho}$. A similar argument holds for $\hat{\alpha}_n^*$ and it holds that $\hat{\alpha}_n^* \leqslant \frac{M}{\rho}$.

Denote $\underline{\alpha} := \alpha^*/2$, $\overline{\alpha} := \frac{M}{\rho}$, and

$$\tau := \min\Big\{\underline{\alpha} \log(\mathbb{E}_{s' \sim P_f}[e^{\frac{-V(s')}{\underline{\alpha}}}]) + \underline{\alpha}\rho, \overline{\alpha} \log(\mathbb{E}_{s' \sim P_f}[e^{\frac{-V(s')}{\overline{\alpha}}}]) + \overline{\alpha}\rho\Big\} - \alpha^* \log\Big(\mathbb{E}_{s' \sim P_f}[e^{\frac{-V(s')}{\alpha^*}}]\Big) - \alpha^*\rho.$$

We first show that,

$$\Big|\log(\frac{\mathbb{E}_{s' \sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}]}{\mathbb{E}_{s' \sim P_f}[e^{-\frac{V(s')}{\alpha}}]})\Big| \leqslant e^{\frac{M}{\alpha}} |\mathbb{E}_{s' \sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}] - \mathbb{E}_{s' \sim P_f}[e^{\frac{-V(s')}{\alpha}}]|. \qquad (70)$$

Consider 2 cases: $\mathbb{E}_{s' \sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}] \geqslant \mathbb{E}_{s' \sim P_f}[e^{\frac{-V(s')}{\alpha}}]$ and $\mathbb{E}_{s' \sim P_f}[e^{\frac{-V(s')}{\alpha}}] > \mathbb{E}_{s' \sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}]$

Case-1: $\mathbb{E}_{s' \sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}] \geqslant \mathbb{E}_{s' \sim P_f}[e^{\frac{-V(s')}{\alpha}}]$:

$$|\log(\frac{\mathbb{E}_{s' \sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}]}{\mathbb{E}_{s' \sim P_f}[e^{-\frac{V(s')}{\alpha}}]})| = \log(\frac{\mathbb{E}_{s' \sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}]}{\mathbb{E}_{s' \sim P_f}[e^{-\frac{V(s')}{\alpha}}]})$$

$$
\begin{aligned}
&= \log(1 + \frac{\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}] - \mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha}}]}{\mathbb{E}_{s'\sim P_f}[e^{-\frac{V(s')}{\alpha}}]}) \\
&\leqslant \frac{\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}] - \mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha}}]}{\mathbb{E}_{s'\sim P_f}[e^{-\frac{V(s')}{\alpha}}]} \\
&\leqslant e^{\frac{M}{\alpha}}(\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}] - \mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha}}]).
\end{aligned}
$$

Case-2: $\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}] < \mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha}}]$:

$$
\begin{aligned}
|\log(\frac{\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}]}{\mathbb{E}_{s'\sim P_f}[e^{-\frac{V(s')}{\alpha}}]})| &= \log(\frac{\mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha}}]}{\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{-\frac{V(s')}{\alpha}}]}) \\
&= \log(1 + \frac{\mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha}}] - \mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}]}{\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{-\frac{V(s')}{\alpha}}]}) \\
&\leqslant \frac{\mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha}}] - \mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}]}{\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{-\frac{V(s')}{\alpha}}]} \\
&\leqslant e^{\frac{M}{\alpha}}(\mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha}}] - \mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}]).
\end{aligned}
$$

Hence, Equation (70) holds. Then, for $\alpha \in [\underline{\alpha}, \overline{\alpha}]$, we have

$$
\begin{aligned}
&|(\alpha \log(\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}]) + \alpha\rho) - (\alpha \log(\mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha}}]) + \alpha\rho)| \qquad (71)\\
&= \alpha|\log(1 + \frac{\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}] - \mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha}}]}{\mathbb{E}_{s'\sim P_f}[e^{-\frac{V(s')}{\alpha}}]})| \\
&\overset{(i)}{\leqslant} \alpha e^{\frac{M}{\alpha}}|\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}] - \mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha}}]| \\
&\overset{(ii)}{\leqslant} \alpha e^{\frac{M}{\alpha}}\|f(s,a) - \hat{f}_n(s,a)\| \quad (\textit{Lemma } 10) \\
&\overset{(iii)}{\leqslant} \mathcal{O}\Big(\alpha e^{\frac{M}{\alpha}}\beta_n(\delta)\sqrt{\frac{2ed^2\Gamma_{nd}}{n}}\Big) \quad (\text{from}\textit{Equation } (26)). \qquad (72)
\end{aligned}
$$

Here (i) holds from Equation (70), (ii) from Lemma 10 and (iii) from Equation (26).

We further show that $\hat{\alpha}^*_n \in [\underline{\alpha}, \overline{\alpha}]$. The first step in achieving that is to restrict $n > N''(\rho, P_f(s,a)) = \mathcal{O}\Big(4\frac{M^2 e^{\frac{2M}{\alpha}}\beta_n^2(\delta)2ed^2\Gamma_{nd}}{(\rho\tau)^2}\Big)$. It implies that if $\mathcal{O}\Big(\alpha e^{\frac{M}{\alpha}}\beta_n(\delta)\sqrt{\frac{2ed^2\Gamma_{nd}}{n}}\Big) < \tau/2$ and for $n > \max_{s,a} N''(\rho, P_f(s,a))$ from Equation (72) with probability at least $1-\delta$, for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have

$$
\max_{\underline{\alpha}, \alpha^*, \overline{\alpha}} |(\alpha \log(\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}]) + \alpha\rho) - (\alpha \log(\mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha}}]) + \alpha\rho)| \leqslant \tau/2. \qquad (73)
$$

It further implies that

$$
\begin{aligned}
&\max_{\alpha\in[\underline{\alpha},\overline{\alpha}]} \{(-\alpha \log(\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho)\} \\
&\overset{(i)}{\geqslant} -\alpha^* \log(\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha^*}}]) - \alpha^*\rho \\
&\overset{(ii)}{\geqslant} -\alpha^* \log(\mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha^*}}]) - \alpha^*\rho - \tau/2 \\
&\overset{(iii)}{\geqslant} \max\{-\underline{\alpha} \log(\mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\underline{\alpha}}}]) - \underline{\alpha}\rho, -\overline{\alpha} \log(\mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\overline{\alpha}}}]) - \overline{\alpha}\rho\} + \tau/2
\end{aligned}
$$

$$\overset{(iv)}{\geqslant} \max\{-\underline{\alpha}\log(\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\underline{\alpha}}}]) - \underline{\alpha}\rho, -\overline{\alpha}\log(\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\overline{\alpha}}}]) - \overline{\alpha}\rho\}. \tag{74}$$

where (i) follows from the fact that $\alpha^* \in [\underline{\alpha}, \overline{\alpha}]$, (ii) follows from Equation (73), (iii) follows from the definition of τ in Equation (70) and (iv) again follows from Equation (73).

Thus $\hat{\alpha}_n^* \in [\underline{\alpha}, \overline{\alpha}]$ follows from Equation (74) and concavity of $-\alpha\log(\mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho$ w.r.t. α. Note that α^* also belongs in this set. We bound $\max_{\alpha\geqslant 0}\{-\alpha\log(\mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho\} - \max_{\alpha\geqslant 0}\{-\alpha\log(\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho\}$ only between $\alpha \in [\underline{\alpha}, \overline{\alpha}]$ instead of all $\alpha > 0$. As a result, it holds that

$$\begin{aligned}
&\left|\max_{\alpha\in[\underline{\alpha},\overline{\alpha}]}\{-\alpha\log(\mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho\} - \max_{\alpha\in[\underline{\alpha},\overline{\alpha}]}\{-\alpha\log(\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho\}\right| \\
&\leqslant \max_{\alpha\in[\underline{\alpha},\overline{\alpha}]}\left|\{-\alpha\log(\mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho\} - \{-\alpha\log(\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho\}\right|. \\
&= \max_{\alpha\in[\underline{\alpha},\overline{\alpha}]}\alpha|\log(1 + \frac{\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}] - \mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha}}]}{\mathbb{E}_{s'\sim P_f}[e^{-\frac{V(s')}{\alpha}}]})| \\
&\leqslant \max_{\alpha\in[\underline{\alpha},\overline{\alpha}]} 2\alpha e^{\frac{M}{\alpha}}|\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}] - \mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha}}]|. \\
&\leqslant 2\frac{M}{\rho}e^{\frac{M}{\underline{\alpha}}}\max_{\alpha\in[\underline{\alpha},\overline{\alpha}]}|\mathbb{E}_{s'\sim P_{\hat{f}_n}}[e^{\frac{-V(s')}{\alpha}}] - \mathbb{E}_{s'\sim P_f}[e^{\frac{-V(s')}{\alpha}}]|,
\end{aligned} \tag{75}$$

where the first inequality follows from Equation (70) and second inequality follows from the bounds of α. Taking a maximum over all (s, a) gets the desired result.

□

Lemma 10. *(Bound by difference between estimated model $\hat{f}_n$ and true f) For any value function $V(s') : \mathcal{S} \to [0, 1/(1-\gamma)]$ and any $\alpha > 0$, it holds that*

$$|\mathbb{E}_{s'\sim P_{\hat{f}_n}(s,a)}[e^{\frac{-V(s')}{\alpha}}] - \mathbb{E}_{s'\sim P_f(s,a)}[e^{\frac{-V(s')}{\alpha}}]| \leqslant \sigma^{-1}\|f(s,a) - \hat{f}_n(s,a)\|,$$

where $P_{\hat{f}_n}(s,a) = \mathcal{N}(\hat{f}_n(s,a), \sigma^2 I)$ and $P_f(s,a) = \mathcal{N}(f(s,a), \sigma^2 I)$.

Proof.

$$\begin{aligned}
\left|\mathbb{E}_{s'\sim P_{\hat{f}_n}(s,a)}[e^{\frac{-V(s')}{\alpha}}] - \mathbb{E}_{s'\sim P_f(s,a)}[e^{\frac{-V(s')}{\alpha}}]\right| &= \left|\int_{\mathbb{R}^d}\frac{1}{\sqrt{(2\pi\sigma^2)^d}}e^{\frac{-V(s')}{\alpha}}(e^{-\frac{\|x-f(s,a)\|^2}{2\sigma^2}} - e^{-\frac{\|x-\hat{f}_n(s,a)\|^2}{2\sigma^2}})\right| \\
&\leqslant \int_{\mathbb{R}^d}\frac{1}{\sqrt{(2\pi\sigma^2)^d}}e^{\frac{-V(s')}{\alpha}}\left|e^{-\frac{\|x-f(s,a)\|^2}{2\sigma^2}} - e^{-\frac{\|x-\hat{f}_n(s,a)\|^2}{2\sigma^2}}\right| \\
&\leqslant \int_{\mathbb{R}^d}\frac{1}{\sqrt{(2\pi\sigma^2)^d}}\left|e^{-\frac{\|x-f(s,a)\|^2}{2\sigma^2}} - e^{-\frac{\|x-\hat{f}_n(s,a)\|^2}{2\sigma^2}}\right| \\
&\overset{(i)}{\leqslant} 2\cdot\mathrm{TV}(P_{\hat{f}_n}(s,a), P_f(s,a)) \\
&\overset{(ii)}{\leqslant} 2\sqrt{\mathrm{KL}(P_{\hat{f}_n}(s,a), P_f(s,a))/2} \\
&\overset{(iii)}{\leqslant} 2\sqrt{\|f(s,a) - \hat{f}_n(s,a)\|^2/4\sigma^2} \\
&\leqslant \|f(s,a) - \hat{f}_n(s,a)\|/\sigma,
\end{aligned}$$

where (i) follows from the definition of Total Variation (TV) distance between any two multivariate Gaussians, (ii) uses the Pinsker's inequality, (iii) uses the formula for KL-divergence between multivariate Gaussian distributions.

□

Lemma 11. *(Proposition-2 in Hu and Hong (2013)) For any function $V(\cdot) : \mathcal{S} \to [0, 1/(1-\gamma)]$ and random variable $Y = V(s')$ for $s' \sim P_f(s,a)$, we have*

$$\lim_{\alpha \to 0} [-\alpha \log(\mathbb{E}_{s' \sim P_f(s,a)}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho] = \text{ESI}(Y),$$

where $\text{ESI}(Y) = \sup\{t \in \mathbb{R} : \mathbb{P}\{Y < t\} = 0\}$ *(essential infimum).*

Proof. Consider the case when $M > \text{ESI}(Y)$. Let $\kappa_M = \mathbb{P}(V(s') \leqslant M) = \int_{s'} \mathbb{1}(V(s') \leqslant M) e^{-\frac{\|s' - f(s,a)\|^2}{\sigma^2}}$. It holds that

$$\begin{aligned}
&- \alpha \log \Big(\mathbb{E}_{s' \sim P_f(s,a)}[e^{\frac{-V(s')}{\alpha}}] \Big) && (76)\\
&= -\alpha \log \Big(\mathbb{E}_{s' \sim P_f(s,a)}[\mathbb{1}(V(s') \leqslant M) e^{\frac{-V(s')}{\alpha}} + \mathbb{1}(V(s') > M) e^{\frac{-V(s')}{\alpha}}] \Big) \\
&\leqslant -\alpha \log \Big(\mathbb{E}_{s' \sim P_f(s,a)}[\mathbb{1}(V(s') \leqslant M) e^{\frac{-V(s')}{\alpha}}] \Big) \\
&\leqslant -\alpha \log \Big(\mathbb{E}_{s' \sim P_f(s,a)}[\mathbb{1}(V(s') \leqslant M) e^{\frac{-M}{\alpha}}] \Big) \\
&\leqslant -\alpha \log \Big(\kappa_M e^{\frac{-M}{\alpha}}] \Big) \\
&= M - \alpha \log(\kappa_M). && (77)
\end{aligned}$$

Thus for any $M > \text{ESI}(Y)$, we have

$$\lim_{\alpha \to 0} [\{-\alpha \log(\mathbb{E}_{s' \sim P_f(s,a)}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho] \leqslant M.$$

Combining with the fact that $\lim_{\alpha \to 0} [\{-\alpha \log(\mathbb{E}_{s' \sim P_f(s,a)}[e^{\frac{-V(s')}{\alpha}}]) - \alpha\rho] \geqslant \text{ESI}(Y)$, we get the desired result. □

Lemma 12. *($\zeta-$cover construction) For $\mathcal{V}$ denoting the set of value functions from $\mathcal{S} \to [0, 1/(1-\gamma)]$, $\overline{\alpha} = M/\rho$, $\underline{\alpha}$ as defined in Lemma 9 we have with probability at least $1-\delta$,*

$$\begin{aligned}
\max_{V \in \mathcal{V}} \max_{s,a} 2\overline{\alpha} e^{\frac{M}{\underline{\alpha}}} \max_{\alpha \in [\underline{\alpha}, \overline{\alpha}]} &|\mathbb{E}_{s' \sim P_{\hat{f}_n}(s,a)}[e^{\frac{-V(s')}{\alpha}}] - \mathbb{E}_{s' \sim P_f(s,a)}[e^{\frac{-V(s')}{\alpha}}]| \\
&\leqslant \mathcal{O}\Big(2(\frac{M}{\rho}) e^{\frac{M}{\alpha_{kl}}} e^{\frac{\zeta}{\alpha_{kl}}} \frac{\beta_n(\delta)\sqrt{2ed^2\Gamma_{nd}}}{\sqrt{n}} \Big).
\end{aligned}$$

Proof. Let $\mathcal{N}_{\mathcal{V}}(\zeta)$ be the $\zeta-$ cover of the set $\mathcal{V}$. By definition, there exists $V' \in \mathcal{N}_{\mathcal{V}}(\zeta)$ such that $\|V' - V\| \leqslant \zeta$ for every $V \in \mathcal{V}$.

$$\begin{aligned}
&|\mathbb{E}_{s' \sim P_{\hat{f}_n}(s,a)}[e^{-V(s')/\alpha}] - \mathbb{E}_{s' \sim P_f(s,a)}[e^{-V(s')/\alpha}]| \\
&\leqslant |\int_{\mathbb{R}^d} \frac{1}{\sqrt{(2\pi\sigma^2)^d}} e^{\frac{-V(s')}{\alpha}} (e^{-\frac{\|s' - f(s,a)\|^2}{\sigma^2}} - e^{-\frac{\|s' - \hat{f}_n(s,a)\|^2}{\sigma^2}})| \\
&\leqslant \int_{\mathbb{R}^d} \frac{1}{\sqrt{(2\pi\sigma^2)^d}} e^{\frac{-V(s')}{\alpha}} |e^{-\frac{\|s' - f(s,a)\|^2}{\sigma^2}} - e^{-\frac{\|s' - \hat{f}_n(s,a)\|^2}{\sigma^2}}| \\
&\leqslant \int_{\mathbb{R}^d} \frac{1}{\sqrt{(2\pi\sigma^2)^d}} e^{\frac{-V(s') + V'(s')}{\alpha}} e^{\frac{-V'(s')}{\alpha}} |e^{-\frac{\|s' - f(s,a)\|^2}{\sigma^2}} - e^{-\frac{\|s' - \hat{f}_n(s,a)\|^2}{\sigma^2}}| \\
&\overset{(i)}{\leqslant} e^{\frac{\zeta}{\alpha_{kl}}} \int_{\mathbb{R}^d} \frac{1}{\sqrt{(2\pi\sigma^2)^d}} e^{\frac{-V'(s')}{\alpha}} |e^{-\frac{\|s' - f(s,a)\|^2}{\sigma^2}} - e^{-\frac{\|s' - \hat{f}_n(s,a)\|^2}{\sigma^2}}| \\
&\leqslant \max_{V' \in \mathcal{N}_{\mathcal{V}}(\zeta)} \max_{s,a} \max_{\alpha \in [\alpha_{kl}, \overline{\alpha}]} e^{\frac{\zeta}{\alpha_{kl}}} \int_{\mathbb{R}^d} \frac{1}{\sqrt{(2\pi\sigma^2)^d}} e^{\frac{-V'(s')}{\alpha}} |e^{-\frac{\|s' - f(s,a)\|^2}{\sigma^2}} - e^{-\frac{\|s' - \hat{f}_n(s,a)\|^2}{\sigma^2}}|. && (78)
\end{aligned}$$

Here (i) is obtained using the fact that $\|V' - V\| \leqslant \zeta$ and α_{kl} is the minimum value of $\underline{\alpha}$ as defined in Lemma 9. Using Equation (78), we bound uniformly over all $V \in \mathcal{V}$, we have

$$\max_{V \in \mathcal{V}} \max_{s,a} 2\overline{\alpha} e^{\frac{M}{\alpha_{kl}}} \max_{\alpha \in [\alpha_{kl}, \overline{\alpha}]} |\mathbb{E}_{s' \sim P_{\hat{f}_n}(s,a)}[e^{\frac{-V(s')}{\alpha}}] - \mathbb{E}_{s' \sim P_f(s,a)}[e^{\frac{-V(s')}{\alpha}}]|$$

$$
\begin{aligned}
&\leqslant \max_{V'\in\mathcal{N}_\mathcal{V}(\zeta)} \max_{s,a} \max_{\alpha\in[\alpha_{kl},\overline{\alpha}]} 2\overline{\alpha} e^{\frac{M}{\alpha_{kl}}} e^{\frac{\zeta}{\alpha_{kl}}} \int_{\mathbb{R}^d} \frac{1}{\sqrt{(2\pi\sigma^2)^d}} e^{\frac{-V'(s')}{\alpha}} |e^{-\frac{\|s'-f(s,a)\|^2}{\sigma^2}} - e^{-\frac{\|s'-\hat{f}_n(s,a)\|^2}{\sigma^2}}| \\
&\leqslant \max_{V'\in\mathcal{N}_\mathcal{V}(\zeta)} \max_{s,a} \max_{\alpha\in[\alpha_{kl},\overline{\alpha}]} 2\overline{\alpha} e^{\frac{M}{\alpha_{kl}}} e^{\frac{\zeta}{\alpha_{kl}}} \int_{\mathbb{R}^d} \frac{1}{\sqrt{(2\pi\sigma^2)^d}} |e^{-\frac{\|s'-f(s,a)\|^2}{\sigma^2}} - e^{-\frac{\|s'-\hat{f}_n(s,a)\|^2}{\sigma^2}}| \qquad (79)\\
&\overset{(i)}{\leqslant} \max_{s,a} 4\overline{\alpha}\sigma^{-1} e^{\frac{M}{\alpha_{kl}}} e^{\frac{\zeta}{\alpha_{kl}}} \|f(s,a) - \hat{f}_n(s,a)\| \\
&\overset{(ii)}{\leqslant} \mathcal{O}\Big(2(\frac{M}{\rho}) e^{\frac{M}{\alpha_{kl}}} e^{\frac{\zeta}{\alpha_{kl}}} \frac{\beta_n(\delta)\sqrt{2ed^2\Gamma_{nd}}}{\sigma\sqrt{n}}\Big)
\end{aligned}
$$

Here (i) follows from Lemma 10 and by the fact that none of the remaining terms inside max depend on V' or α. And (ii) follows from $\overline{\alpha} = \frac{M}{\rho}$ and Equation (26).

□

C Other Uncertainty Sets

C.1 Chi-Square Uncertainty Set

The f-divergence (Ali and Silvey (1966); Csiszár (1967)) between probability measures P and P_0 defined over $\mathcal{X}$ for a convex function $f : \mathbb{R} \to \bar{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{\infty\}$ satisfying $f(1) = 0$ and $f(t) = \infty$ for any $t < 0$ is defined as follows:

$$D_f(P||P_0) = \int f\Big(\frac{dP}{dP_0}\Big) dP_0. \qquad (80)$$

Specifically Duchi and Namkoong (2021) considers the Cressie-Read family of f-divergences (Cressie and Read (1984), see Appendix C.1) which includes χ^2 divergence ($k = 2$), etc. This family of f-divergences can be parametrized by $k \in (-\infty, \infty)\backslash\{0, 1\}$ with $f_k(t) := \frac{t^k - kt + k - 1}{k(k-1)}$ Using this, we state the reformulation result from Duchi and Namkoong (2021, Lemma-1).

Lemma 13. *For $k \in (1, \infty)$, $k_* = k/k - 1$, any $\rho > 0$ and $c_k(\rho) = (1 + k(k-1)\rho)^{\frac{1}{k}}$ and $X \sim P_0$ where P_0 is any probability distribution over $\mathcal{X}$ with $H : \mathcal{X} \to \mathbb{R}$, we have*

$$\sup_{P: D_{f_k}(P||P_0)\leqslant\rho} \mathbb{E}_P[H(X)] = \inf_{\eta\in\mathbb{R}} \{c_k(\rho)(\mathbb{E}_{P_0}[(H(X) - \eta)_+^{k_*}])^{\frac{1}{k_*}} + \eta\}. \qquad (81)$$

Theorem 3. *(Sample Complexity under χ^2 uncertainty set) Consider a robust MDP (see Section 2) with nominal transition dynamics f and uncertainty set defined as in Equation (2) w.r.t. χ^2 divergence. For π^* denoting the robust optimal policy w.r.t. nominal transition dynamics f and π_N^* denoting the robust optimal policy w.r.t. learned nominal transition dynamics $\hat{f}_N$ via Algorithm 1, and $\delta \in (0, 1)$, $\epsilon \in (0, \frac{1}{1-\gamma})$,it holds that $\max_s |V^R_{\pi_N^*, f}(s) - V^R_{\pi^*, f}(s)| \leqslant \epsilon$ with probability at least $1 - \delta$ for any $N \geqslant N_{\chi^2}$, where*

$$N_{\chi^2} = \mathcal{O}\Big(\Big(\frac{1 + 2\rho}{\sqrt{1 + 2\rho} - 1}\Big)^4 \frac{\gamma^4 \beta_n(\delta)^2 d^2 \gamma_{nd}}{\epsilon^4 (1 - \gamma)^8}\Big). \qquad (82)$$

Proof. **Step (i):** As detailed in the proof outline of Section 4, in order to bound $V^{\mathrm{R}}_{\hat{\pi}_n, f}(s) - V^{\mathrm{R}}_{\pi^*, f}(s)$, we begin by adding and subtracting $V^{\mathrm{R}}_{\hat{\pi}_n, \hat{f}_n}(s)$ which is the robust value function w.r.t. the nominal transition dynamics $\hat{f}_n$ and its corresponding optimal policy $\hat{\pi}_n$. Then, we split the difference into two terms as follows:

$$V^{\mathrm{R}}_{\hat{\pi}_n, f}(s) - V^{\mathrm{R}}_{\pi^*, f}(s) = \underbrace{V^{\mathrm{R}}_{\hat{\pi}_n, f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n, \hat{f}_n}(s)}_{(i)} + \underbrace{V^{\mathrm{R}}_{\hat{\pi}_n, \hat{f}_n}(s) - V^{\mathrm{R}}_{\pi^*, f}(s)}_{(ii)}. \qquad (83)$$

In order to not disturb the flow of the proof we bound (i) and (ii) separately Lemma 6 and Lemma 7 respectively. From Lemma 6, we obtain that

$$(i) \leqslant \max_s \Big| V^{\mathrm{R}}_{\hat{\pi}_n, f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n, \hat{f}_n}(s) \Big|$$

$$\leqslant \frac{\gamma}{1-\gamma}\max_s \Big| \inf_{\chi^2(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big] - \inf_{\chi^2(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big]\Big|. \tag{84}$$

And from Lemma 7, we obtain that

$$\begin{aligned}(ii) &\leqslant \max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)\Big| \\ &\leqslant \frac{\gamma}{1-\gamma}\max_s \Big| \inf_{\chi^2(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\pi^*,f}(s')\Big] - \inf_{\chi^2(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\pi^*,f}(s')\Big]\Big|. \end{aligned} \tag{85}$$

Note that both these terms in Equations (84) and (85) are of the form mentioned in the **Step (i)** of Section 4.

Step (ii): Next, corresponding to **Step (ii)** of the proof outline in Section 4, we use Lemma 13 to bound Equations (84) and (85). Denote $M := \frac{1}{1-\gamma} \geqslant \max_s V^{\mathrm{R}}_{\pi}(s)$ and $c_2(\rho) := \sqrt{1+2\rho}$ for convenience. Using Equation (84) and Lemma 14 (internally using Lemma 13), it holds that

$$\begin{aligned}(i) &\leqslant \max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s)\Big| \\ &\leqslant \frac{1}{1-\gamma}\max_s \Big|\gamma \inf_{\chi^2(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big] - \gamma \inf_{\chi^2(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big]\Big| \\ &\leqslant \max_{s,a}\left(\frac{\gamma\sqrt{1+2\rho}}{1-\gamma}\sup_{\eta\in[0,\frac{c_2(\rho)M}{c_2(\rho)-1}]}\Big\{\Big|\mathbb{E}_{P_f(s,a)}[(-V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')+\eta)^2_+] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')+\eta)^2_+]\Big|^{\frac{1}{2}}\Big\}\right). \qquad (86) \\ &\leqslant \max_{V(\cdot)\in\mathcal{V}}\max_{s,a}\left(\frac{\gamma\sqrt{1+2\rho}}{1-\gamma}\sup_{\eta\in[0,\frac{c_2(\rho)M}{c_2(\rho)-1}]}\Big\{\Big|\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)^2_+] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')+\eta)^2_+]\Big|^{\frac{1}{2}}\Big\}\right). \qquad (87)\end{aligned}$$

We can bound (ii) similarly.

$$\begin{aligned}(ii) &\leqslant \max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)\Big| \qquad (88) \\ &\leqslant \max_{V(\cdot)\in\mathcal{V}}\max_{s,a}\left(\frac{\gamma\sqrt{1+2\rho}}{1-\gamma}\sup_{\eta\in[0,\frac{c_2(\rho)M}{c_2(\rho)-1}]}\Big\{\Big|\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)^2_+] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')+\eta)^2_+]\Big|^{\frac{1}{2}}\Big\}\right). \qquad (89)\end{aligned}$$

Step (iii): Next, we want to utilize the learning error bound (Equation (26)) that bounds the difference between the means of true nominal transition dynamics P_f and learned nominal transition dynamics $P_{\hat{f}_n}$ to bound Equations (87) and (89).

We begin by bounding the difference $\Big|\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)^2_+] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')+\eta)^2_+]\Big|$, by the difference in means of P_f and $P_{\hat{f}_n}$ in Lemma 17. Since Equation (87) has a max over all value functions, we introduce a covering number argument in Lemma 15 to reform it to a max over the functions in the $\zeta-$covering set. We then use Lemma 17 to obtain bounds in terms of maximum information gain Γ_{Nd} (Equation (9)) and ζ. Further details regarding the covering number argument are deferred to Lemma 15. Then, we apply the result of Lemma 15 with $\zeta = 1$ (defined in Lemma 15) on Equation (87). Then, it holds that

$$(i) \leqslant \max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s)\Big| = \mathcal{O}\left(\Big(\frac{\gamma(c_2(\rho))^2M^2}{c_2(\rho)-1}\Big)\Big(\frac{\beta_n(\delta)\sqrt{2ed^2\gamma_{nd}}}{\sigma\sqrt{n}}\Big)^{\frac{1}{2}}\right). \tag{90}$$

Note that β_n, which appears in Lemma 3, has a logarithmic dependence on n. Similarly, from Equation (89), and Lemmas 15 and 17, we obtain

$$(ii) \leqslant \max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)\Big| = \mathcal{O}\left(\Big(\frac{\gamma(c_2(\rho))^2M^2}{c_2(\rho)-1}\Big)\Big(\frac{\beta_n(\delta)\sqrt{2ed^2\gamma_{nd}}}{\sigma\sqrt{n}}\Big)^{\frac{1}{2}}\right). \tag{91}$$

Note that we want to bound $V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\pi^*,f}(s) = (i) + (ii)$ over all $s \in \mathcal{S}$. Using $\max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)\Big| \leqslant \max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)\Big| + \max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) - V^{\mathrm{R}}_{\pi^*_n,f}(s)\Big|$ and substituting M by $1/(1-\gamma)$, we obtain from

Equation (90) and Equation (91)

$$\max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)\Big| = \mathcal{O}\bigg(\Big(\frac{\gamma(c_2(\rho))^2M^2}{c_2(\rho)-1}\Big)\Big(\frac{\beta_n(\delta)\sqrt{2ed^2\gamma_{nd}}}{\sigma\sqrt{n}}\Big)^{\frac{1}{2}}\bigg).$$

Finally, to ensure that $\max_s |V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)| \leqslant \epsilon$, it suffices to have

$$\max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)\Big| = \mathcal{O}\bigg(\Big(\frac{\gamma(c_2(\rho))^2M^2}{c_2(\rho)-1}\Big)\Big(\frac{\beta_n(\delta)\sqrt{2ed^2\gamma_{nd}}}{\sigma\sqrt{n}}\Big)^{\frac{1}{2}}\bigg) = \epsilon.$$

Moving $\sqrt{n}$ and ϵ to opposite sides and squaring both sides twice, we obtain

$$n = \mathcal{O}\Big(\Big(\frac{1+2\rho}{\sqrt{1+2\rho}-1}\Big)^4\frac{\gamma^4\beta_n(\delta)^2d^2\gamma_{nd}}{\sigma^2\epsilon^4(1-\gamma)^8}\Big).$$

□

Lemma 14. *(Simplification using Lemma 13 reformulation) For any value function V from $\mathcal{S}\to[0,1/(1-\gamma)]$, it holds that*

$$\max_s \Big|\inf_{\chi^2(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho}\mathbb{E}_{s'\sim p}\Big[V(s')\Big] - \inf_{\chi^2(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho}\mathbb{E}_{s'\sim p}\Big[V(s')\Big]\Big| \leqslant$$
$$\max_{s,a} c_2(\rho)\sup_{\eta\in[0,\frac{c_2(\rho)M}{c_2(\rho)-1}]}\{|\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)_+^2] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')+\eta)_+^2]|^{\frac{1}{2}}\}, \quad (92)$$

where $c_2(\rho)=\sqrt{1+2\rho}$ and $M=1/(1-\gamma)$.

Proof. First note that,

$$\max_s |\inf_{\chi^2(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho}\mathbb{E}_{s'\sim p}\Big[V(s')\Big] - \inf_{\chi^2(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho}\mathbb{E}_{s'\sim p}\Big[V(s')\Big]| \leqslant$$
$$\max_{s,a}\Big|\inf_{\chi^2(p||P_{\hat{f}_n}(s,a))\leqslant\rho}\mathbb{E}_{s'\sim p}\Big[V(s')\Big] - \inf_{\chi^2(p||P_f(s,a))\leqslant\rho}\mathbb{E}_{s'\sim p}\Big[V(s')\Big]\Big|. \quad (93)$$

Using Lemma 13 and focusing to bound right side of Equation (93) for one particular (s,a) state-action pair, we obtain

$$\Big|\inf_{\chi^2(p||P_{\hat{f}_n}(s,a))\leqslant\rho}\mathbb{E}_{s'\sim p}\Big[V(s')\Big] - \inf_{\chi^2(p||P_f(s,a))\leqslant\rho}\mathbb{E}_{s'\sim p}\Big[V(s')\Big]\Big| =$$
$$\Big|\sup_{\eta\in\mathbb{R}}\{-c_2(\rho)(\mathbb{E}_{P_f(s,a)}[(-V(s')-\eta)_+^2])^{\frac{1}{2}}-\eta\} - \sup_{\eta\in\mathbb{R}}\{-c_2(\rho)(\mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')-\eta)_+^2])^{\frac{1}{2}}-\eta\}\Big|$$
$$\overset{(i)}{=}\Big|\sup_{\eta\in\mathbb{R}}\{-c_2(\rho)(\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)_+^2])^{\frac{1}{2}}+\eta\} - \sup_{\eta\in\mathbb{R}}\{-c_2(\rho)(\mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')+\eta)_+^2])^{\frac{1}{2}}+\eta\}\Big|, \quad (94)$$

where (i) is obtained by replacing η with $-\eta$.
Let $g_{\chi^2}(\eta, P_f(s,a)) := \Big(-c_2(\rho)(\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)_+^2])^{\frac{1}{2}}+\eta\Big)$. Note that $g_{\chi^2}(\eta,P_f(s,a))$ satisfies the following: For $\eta\leqslant 0$ (implying $(-V(s')+\eta)\leqslant 0$ and $(-V(s')+\eta)_+=0$),

$$g_{\chi^2}(\eta,P_f(s,a)) = \eta \leqslant 0. \quad (95)$$

And for $\eta=\frac{c_2(\rho)M}{c_2(\rho)-1}>0$,

$$g_{\chi}(\tfrac{c_2(\rho)M}{c_2(\rho)-1},P_f(s,a)) = -c_2(\rho)(\mathbb{E}_{P_f(s,a)}[(-V(s')+\tfrac{c_2(\rho)M}{c_2(\rho)-1})_+^2])^{\frac{1}{2}} + \tfrac{c_2(\rho)M}{c_2(\rho)-1}$$
$$\overset{(i)}{\leqslant}\tfrac{c_2(\rho)M}{c_2(\rho)-1} - c_2(\rho)(\mathbb{E}_{P_f(s,a)}[(-M+\tfrac{c_2(\rho)M}{c_2(\rho)-1})_+^2])^{\frac{1}{2}}$$

$$
\begin{aligned}
&\leqslant \tfrac{c_2(\rho)M}{c_2(\rho)-1} - c_2(\rho)(\mathbb{E}_{P_f(s,a)}[(\tfrac{M}{c_2(\rho)-1})^2_+])^{\frac{1}{2}} \\
&\leqslant \tfrac{c_2(\rho)M}{c_2(\rho)-1} - \tfrac{c_2(\rho)M}{c_2(\rho)-1} \\
&= 0, \qquad (96)
\end{aligned}
$$

where (i) follows from the fact that the random variable $V(s')$ is bounded by $M = 1/1-\gamma$. A similar result can be shown for $g_{\chi^2}(\eta, P_{\hat{f}_n}(s,a))$ (or for any P). Along with the convexity of $\eta \to g_\chi(\eta, P)$ (Duchi and Namkoong (2021)), and $\inf_{\chi^2(p||P)\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V(s')\Big] \geqslant 0$, Equation (95) and Equation (96) imply that the sup is attained between $[0, \frac{c_2(\rho)M}{c_2(\rho)-1}]$ for both $\sup_{\eta\in\mathbb{R}} g_\chi(\eta, P_f(s,a))$ and $\sup_{\eta\in\mathbb{R}} g_\chi(\eta, P_{\hat{f}_n}(s,a))$. Using this in Equation (94) we have,

$$
\begin{aligned}
&\Big|\sup_{\eta\in\mathbb{R}}\{g_\chi(\eta, P_f(s,a))\} - \sup_{\eta\in\mathbb{R}}\{g_\chi(\eta, P_{\hat{f}_n}(s,a))\}\Big| && (97)\\
&= \Big|\sup_{\eta\in[0,\frac{c_2(\rho)M}{c_2(\rho)-1}]}\{g(_\chi\eta, P_f(s,a))\} - \sup_{\eta\in[0,\frac{c_2(\rho)M}{c_2(\rho)-1}]}\{g_\chi(\eta, P_{\hat{f}_n}(s,a))\}\}\Big| && (98)\\
&\leqslant \sup_{\eta\in[0,\frac{c_2(\rho)M}{c_2(\rho)-1}]}\{|g_\chi(\eta, P_f(s,a)) - g_\chi(\eta, P_{\hat{f}_n}(s,a))|\} && (99)\\
&\leqslant \sup_{\eta\in[0,\frac{c_2(\rho)M}{c_2(\rho)-1}]}\{|c_2(\rho)(\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)^2_+])^{\frac{1}{2}} - c_2(\rho)\mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')+\eta)^2_+])^{\frac{1}{2}}|\} && (100)\\
&\leqslant c_2(\rho)\sup_{\eta\in[0,\frac{c_2(\rho)M}{c_2(\rho)-1}]}\{|\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)^2_+] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')+\eta)^2_+]|^{\frac{1}{2}}\}. && (101)
\end{aligned}
$$

The last step is obtained using the basic inequality $|\sqrt{a}-\sqrt{b}| \leqslant \sqrt{|a-b|}$.

□

Lemma 15. *($\zeta-$cover construction) For $\mathcal{V}$ denoting the set of value functions from $\mathcal{S} \to [0, 1/(1-\gamma)]$ it holds with probability at least $1-\delta$,*

$$
\max_{V\in\mathcal{V}}\max_{s,a}\sup_{\eta\in[0,\frac{c_2(\rho)M}{c_2(\rho)-1}]}\{|\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)^2_+] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')+\eta)^2_+]|^{\frac{1}{2}}\} \leqslant \mathcal{O}\left(\Big(\frac{c_2(\rho)M}{c_2(\rho)-1}\Big)\Big(\frac{\beta_n(\delta)\sqrt{2ed^2\gamma_{nd}}}{\sigma\sqrt{n}}\Big)^{\frac{1}{2}}\right), \quad (102)
$$

where $c_2(\rho) = \sqrt{1+2\rho}$, $M = 1/(1-\gamma)$.

Proof. Let $\mathcal{N}_\mathcal{V}(\zeta)$ be the $\zeta-$ cover of the set $\mathcal{V}$. By definition, there exists $V' \in \mathcal{N}_\mathcal{V}(\zeta)$ such that $\|V'-V\| \leqslant \zeta$ for every $V \in \mathcal{V}$.

$$
\begin{aligned}
&|\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)^2_+] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')+\eta)^2_+]| \\
&\leqslant |\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)^2_+] - \mathbb{E}_{P_f(s,a)}[(-V'(s')+\eta)^2_+]| \\
&\quad + |\mathbb{E}_{P_f(s,a)}[(-V'(s')+\eta)^2_+] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V'(s')+\eta)^2_+]| \\
&\quad + |\mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V'(s')+\eta)^2_+] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')+\eta)^2_+]|. \qquad (103)\\
&\overset{(i)}{\leqslant} 4\|V'-V\|^2 + 4\eta\|V'-V\| + |\mathbb{E}_{P_f(s,a)}[(-V'(s')+\eta)^2_+] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V'(s')+\eta)^2_+]|, \qquad (104)
\end{aligned}
$$

where (i) follows from Lemma 16. Using Equation (104) we bound uniformly over all $V \in \mathcal{V}$,

$$
\max_{V\in\mathcal{V}}\max_{s,a}\sup_{\eta\in[0,\frac{c_2(\rho)M}{c_2(\rho)-1}]}\{|\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)^2_+] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')+\eta)^2_+]|^{\frac{1}{2}}\} \quad (105)
$$

$$
\begin{aligned}
&\leqslant \max_{V'\in\mathcal{N}_{\mathcal{V}}(\zeta)} \max_{s,a} \sup_{\eta\in[0,\frac{c_2(\rho)M}{c_2(\rho)-1}]} \Bigg\{ \Big(4\|V'-V\|^2 + 4\eta\|V'-V\| + |\mathbb{E}_{P_f(s,a)}[(-V'(s')+\eta)_+^2] \\
&\quad - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V'(s')+\eta)_+^2]| \Big)^{\frac{1}{2}} \Bigg\} \\
&\overset{(ii)}{\leqslant} \max_{V'\in\mathcal{N}_{\mathcal{V}}(\zeta)} \max_{s,a} \sup_{\eta\in[0,\frac{c_2(\rho)M}{c_2(\rho)-1}]} \Bigg\{ \Big(\mathbb{E}_{P_f(s,a)}[(-V'(s')+\eta)_+^2] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V'(s')+\eta)_+^2] \Big)^{\frac{1}{2}} \Bigg\} \\
&\quad + \sqrt{4\zeta^2 + 4\zeta\tfrac{c_2(\rho)M}{c_2(\rho)-1}} \\
&\overset{(iii)}{\leqslant} \max_{V'\in\mathcal{N}_{\mathcal{V}}(\zeta)} \max_{s,a} \sup_{\eta\in[0,\frac{c_2(\rho)M}{c_2(\rho)-1}]} \left\{ \left(\tfrac{c_2(\rho)M}{c_2(\rho)-1}\right) \sqrt{2\sigma^{-1}\|f(s,a)-\hat{f}_n(s,a)\|} \right\} + \sqrt{4\zeta^2 + 4\zeta\tfrac{c_2(\rho)M}{c_2(\rho)-1}} \\
&\overset{(iv)}{\leqslant} \mathcal{O}\left(\left(\tfrac{c_2(\rho)M}{c_2(\rho)-1}\right) \left(\tfrac{\beta_n(\delta)\sqrt{2ed^2\gamma_{nd}}}{\sigma\sqrt{n}}\right)^{\frac{1}{2}} \right) + \sqrt{4\zeta^2 + 4\zeta\tfrac{c_2(\rho)M}{c_2(\rho)-1}} \qquad (106) \\
&\overset{(v)}{\leqslant} \mathcal{O}\left(\left(\tfrac{c_2(\rho)M}{c_2(\rho)-1}\right) \left(\tfrac{\beta_n(\delta)\sqrt{2ed^2\gamma_{nd}}}{\sigma\sqrt{n}}\right)^{\frac{1}{2}} \right), \qquad (107)
\end{aligned}
$$

where (ii) follows from $\|V'-V\| \leqslant \zeta$ and $\eta \leqslant \frac{c_2(\rho)M}{c_2(\rho)-1}$, (iii) follows from Lemma 17, (iv) follows from Equation (26), and (v) follows from substituing $\zeta = 1$ (or any constant).

□

Lemma 16. *For any two value functions V, V' from $\mathcal{S} \to [0, 1/(1-\gamma)]$, it holds that*

$$
\Big| \mathbb{E}_{P_f(s,a)}[(-V'(s')+\eta)_+^2] - \mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)_+^2] \Big| \leqslant 2\|V'-V\|^2 + 2\eta\|V'-V\|. \qquad (108)
$$

Proof. Let $p_{P_f(s,a)}(\cdot)$ denote the probability density function of $P_f(s,a)$. Then,

$$
\begin{aligned}
&\mathbb{E}_{P_f(s,a)}[(-V'(s')+\eta)_+^2] - \mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)_+^2] \\
&\leqslant \int_{s'\sim P_f(s,a)} \Big(\mathbb{1}(V'(s')<\eta)(-V'(s')+\eta)^2 - \mathbb{1}(V(s')<\eta)(-V(s')+\eta)^2 \Big) p_{P_f(s,a)}(s')ds'. \\
&\leqslant \underbrace{\int_{s'\sim P_f(s,a)} \Big(\mathbb{1}(V'(s')<\eta) - \mathbb{1}(V(s')<\eta) \Big)(-V'(s')+\eta)^2 p_{P_f(s,a)}(s')ds'}_{(i)} \\
&\quad + \underbrace{\int_{s'\sim P_f(s,a)} \mathbb{1}(V(s')<\eta) \Big((-V'(s')+\eta)^2 - (-V(s')+\eta)^2 \Big) p_{P_f(s,a)}(s')ds'}_{(ii)}.
\end{aligned} \qquad (109)
$$

where the last inequality is obtained by adding and subtracting $\mathbb{1}(V(s')<\eta)(-V'(s')+\eta)^2$.

We begin by bounding (ii). We have,

$$
\begin{aligned}
(ii) &= \int_{s'\sim P_f(s,a)} \mathbb{1}(V(s')<\eta) \Big((-V'(s')+\eta)^2 - (-V(s')+\eta)^2 \Big) p_{P_f(s,a)}(s')ds' \\
&= \int_{s'\sim P_f(s,a)} \mathbb{1}(V(s')<\eta) \Big(-V'(s') + V(s') \Big) \Big(-V'(s') - V(s') + 2\eta \Big) p_{P_f(s,a)}(s')ds' \\
&\leqslant \int_{s'\sim P_f(s,a)} \mathbb{1}(V(s')<\eta) \Big(\mathbb{1}(V'(s')<\eta) + \mathbb{1}(V'(s')\geqslant\eta) \Big) \Big(-V'(s') + V(s') \Big) \\
&\qquad \Big(-V'(s') - V(s') + 2\eta \Big) p_{P_f(s,a)}(s')ds'
\end{aligned}
$$

$$\leqslant \underbrace{\int \mathbb{1}(V(s'), V'(s') < \eta)(-V'(s') + V(s'))(-V'(s') - V(s') + 2\eta) p_{P_f(s,a)}(s') ds'}_{(ii-a)} + \underbrace{\int \mathbb{1}(V(s') < \eta \leqslant V'(s'))(-V'(s') + V(s'))(-V'(s') - V(s') + 2\eta) p_{P_f(s,a)}(s') ds'}_{(ii-b)}. \tag{110}$$

Bounding $(ii-a)$ first, we have,

$$\begin{aligned}
(ii-a) &= \int \mathbb{1}(V(s'), V'(s') < \eta)(-V'(s') + V(s'))(-V'(s') - V(s') + 2\eta) p_{P_f(s,a)}(s') ds' \\
&\overset{(a)}{\leqslant} \int \mathbb{1}(V(s'), V'(s') < \eta) \Big| - V'(s') + V(s') \Big| (-V'(s') - V(s') + 2\eta) p_{P_f(s,a)}(s') ds' \\
&\overset{(b)}{\leqslant} \int\limits_{s' \sim P_f(s,a)} \mathbb{1}(V(s'), V'(s') < \eta) \Big| - V'(s') + V(s') \Big| \Big(2\eta\Big) p_{P_f(s,a)}(s') ds' \\
&\leqslant 2\eta \|V' - V\|,
\end{aligned} \tag{111}$$

where (a) and (b) follows from $(-V'(s') - V(s') + 2\eta) > 0$ as $V(s'), V'(s') < \eta$. And $(ii-b)$ can be bounded as,

$$\begin{aligned}
(ii-b) &= \int \mathbb{1}(V(s') < \eta \leqslant V'(s'))(-V'(s') + V(s'))(-V'(s') - V(s') + 2\eta) p_{P_f(s,a)}(s') ds' \\
&\leqslant \int \mathbb{1}(V(s') < \eta \leqslant V'(s')) \Big| - V'(s') + V(s') \Big| \Big| - V'(s') - V(s') + 2\eta \Big| p_{P_f(s,a)}(s') ds' \\
&\overset{(c)}{\leqslant} \int \mathbb{1}(V(s') < \eta \leqslant V'(s')) \Big| - V'(s') + V(s') \Big| \Big| - V(s') + V'(s') \Big| p_{P_f(s,a)}(s') ds' \\
&\leqslant \int\limits_{s' \sim P_f(s,a)} \mathbb{1}(V(s') < \eta \leqslant V'(s')) \Big| - V'(s') + V(s') \Big|^2 p_{P_f(s,a)}(s') ds' \\
&\leqslant \|V' - V\|^2,
\end{aligned} \tag{112}$$

where (c) follows from $\eta \leqslant V'(s')$. Bounding (i) similarly,

$$\begin{aligned}
i &= \int_{s' \sim P_f(s,a)} \Big(\mathbb{1}(V'(s') < \eta) - \mathbb{1}(V(s') < \eta) \Big) (-V'(s') + \eta)^2 p_{P_f(s,a)}(s') ds' \\
&\leqslant \int_{s' \sim P_f(s,a)} \Big(\mathbb{1}(V'(s') < \eta \leqslant V(s')) \Big) (-V'(s') + \eta)^2 p_{P_f(s,a)}(s') ds' \\
&\leqslant \int_{s' \sim P_f(s,a)} \Big(\mathbb{1}(V'(s') < \eta \leqslant V(s')) \Big) (-V'(s') + V(s'))^2 p_{P_f(s,a)}(s') ds' \\
&\leqslant \|V' - V\|^2.
\end{aligned} \tag{113}$$

Using Equations (109) to (113) we get the desired result. □

Lemma 17. *(Bound by difference between estimated model $\hat{f}_n$ and true f) For any value function $V(s') : \mathcal{S} \to [0, 1/(1-\gamma)]$ and any $\alpha > 0$, it holds that*

$$|\mathbb{E}_{P_f(s,a)}[(-V(s') + \eta)_+^2] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s') + \eta)_+^2]| \leqslant 2\sigma^{-1} \Big(\frac{c_2(\rho) M}{c_2(\rho) - 1} \Big)^2 \|f(s,a) - \hat{f}_n(s,a)\|,$$

where $P_{\hat{f}_n}(s,a) = \mathcal{N}(\hat{f}_n(s,a), \sigma^2 I)$ and $P_f(s,a) = \mathcal{N}(f(s,a), \sigma^2 I)$, $\eta \in [0, \frac{c_2(\rho) M}{c_2(\rho) - 1}]$, $c_2(\rho) = \sqrt{1 + 2\rho}$ and $M = 1/(1-\gamma)$.

Proof.

$$\Big| \mathbb{E}_{P_f(s,a)}[(-V(s') + \eta)_+^2] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s') + \eta)_+^2] \Big|$$

$$
\begin{aligned}
&= \Big| \int_{\mathbb{R}^d} \frac{1}{\sqrt{(2\pi\sigma^2)^d}} (-V(s') + \eta)_+^2 (e^{-\frac{\|x - f(s,a)\|^2}{2\sigma^2}} - e^{-\frac{\|x - \hat{f}_n(s,a)\|^2}{2\sigma^2}}) \Big| \\
&\leqslant \int_{\mathbb{R}^d} \frac{1}{\sqrt{(2\pi\sigma^2)^d}} (-V(s') + \eta)_+^2 \Big| e^{-\frac{\|x - f(s,a)\|^2}{2\sigma^2}} - e^{-\frac{\|x - \hat{f}_n(s,a)\|^2}{2\sigma^2}} \Big| \\
&\overset{(i)}{\leqslant} \Big(\frac{c_2(\rho) M}{c_2(\rho) - 1} \Big)^2 \int_{\mathbb{R}^d} \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \Big| e^{-\frac{\|x - f(s,a)\|^2}{2\sigma^2}} - e^{-\frac{\|x - \hat{f}_n(s,a)\|^2}{2\sigma^2}} \Big| \\
&\overset{(ii)}{\leqslant} 2 \Big(\frac{c_2(\rho) M}{c_2(\rho) - 1} \Big)^2 \cdot \mathrm{TV}(P_{\hat{f}_n}(s,a), P_f(s,a)) \\
&\overset{(iii)}{\leqslant} 2 \Big(\frac{c_2(\rho) M}{c_2(\rho) - 1} \Big)^2 \sqrt{\mathrm{KL}(P_{\hat{f}_n}(s,a), P_f(s,a))/2} \\
&\overset{(iv)}{\leqslant} 2 \Big(\frac{c_2(\rho) M}{c_2(\rho) - 1} \Big)^2 \sqrt{\|f(s,a) - \hat{f}_n(s,a)\|^2 / 4\sigma^2} \\
&\leqslant \Big(\frac{c_2(\rho) M}{c_2(\rho) - 1} \Big)^2 \|f(s,a) - \hat{f}_n(s,a)\| / \sigma,
\end{aligned}
$$

where (i) follows from $(-V(s') + \eta)_+^2 \leqslant \left(\frac{c_2(\rho)M}{c_2(\rho)-1}\right)^2$ as $\eta \leqslant \left(\frac{c_2(\rho)M}{c_2(\rho)-1}\right)$, (ii) follows from the definition of Total Variation (TV) distance between any two multivariate Gaussians, (iii) uses the Pinsker's inequality, and (iv) uses the formula for KL-divergence between multivariate Gaussian distributions.

□

C.2 Total Variation Distance

Similar to lemma 13, we want a similar convex reformulation for the variation distance. We derive such a reformulation starting from the dual reformulation from Shapiro (2017) and Ben-Tal et al. (2013) stated as Proposition-1 in Duchi and Namkoong (2021).

Lemma 18. *For $X \sim P_0$ where P_0 is any probability distribution over $\mathcal{X}$ with $H : \mathcal{X} \to \mathbb{R}$, $\rho > 0$ and, $D_f(P||P_0)$ defined as in Equation* (80) *, it holds that*

$$
\sup_{P : D_f(P||P_0) \leqslant \rho} \mathbb{E}_P[H(X)] = \inf_{\lambda \geqslant 0, \eta \in \mathbb{R}} \Big\{ \mathbb{E}_{P_0}\Big[\lambda f^*\Big(\frac{H(X) - \eta}{\lambda} \Big) \Big] + \lambda\rho + \eta \Big\}. \tag{114}
$$

Note that the total variation distance between two probability distributions P and P_0 is attained by substituting $f_{\mathrm{TV}}(t) = |t-1|$ in $D_f(P||P_0) = \int f\Big(\frac{dP}{dP_0}\Big) dP_0$. The corresponding Fenchel conjugate $f^*_{\mathrm{TV}}(s)$ for $f_{\mathrm{TV}}(t) = |t-1|$ would be

$$
f^*_{\mathrm{TV}}(s) = \begin{cases} -1, & s \leqslant -1 \\ s, & s \in [-1, 1] \\ \infty, & s > 1 \end{cases} \tag{115}
$$

As we require $\inf_{P : \mathrm{TV}(P||P_0) \leqslant \rho} \mathbb{E}_P[H(X)]$, using Equation (114) and replacing η with $-\eta$, we have

$$
\inf_{P : \mathrm{TV}(P||P_0) \leqslant \rho} \mathbb{E}_P[H(X)] = \sup_{\lambda \geqslant 0, \eta \in \mathbb{R}} \{ -\mathbb{E}_{P_0}\Big[\lambda f^*_{\mathrm{TV}}\Big(\frac{-H(X) + \eta}{\lambda} \Big) \Big] - \lambda\rho + \eta \}. \tag{116}
$$

Using Equation (116), we derive a convex reformulation in Lemma 19

Lemma 19. *(Reformulation for total variation distance based on Yang et al. (2022)) For $\rho > 0$ and $X \sim P_0$ where P_0 is any probability distribution over $\mathcal{X}$ with $H : \mathcal{X} \to \mathbb{R}$, for $0 \leqslant H(x) \leqslant \frac{1}{1-\gamma}$ and $ESI(Y) = \sup\{t \in \mathbb{R} : \mathbb{P}\{Y < t\} = 0\}$ (essential infimum), it holds that*

$$
\inf_{P : \mathrm{TV}(P||P_0) \leqslant \rho} \mathbb{E}_P[H(X)] = \sup_{\eta \in [0, \frac{(2+\rho)}{\rho(1-\gamma)}]} \Big\{ -\mathbb{E}_{P_0}[-H(X) + \eta]_+ - \frac{(-ESI(H(x)) + \eta)_+}{2} \rho + \eta \Big\}. \tag{117}
$$

where TV denotes the total variation distance.

Proof. Substituting Equation (115) in Equation (116) to obtain the reformulation for total variation distance, we have

$$\inf_{P:\mathrm{TV}(P||P_0)\leqslant\rho} \mathbb{E}_P[H(X)] \tag{118}$$

$$= \sup_{\lambda\geqslant 0,\eta\in\mathbb{R},\frac{-H(x)+\eta}{\lambda}\leqslant 1} \{-\mathbb{E}_{P_0}\Big[\lambda \max\Big\{\frac{-H(X)+\eta}{\lambda}, -1\Big\}\Big] - \lambda\rho + \eta\} \tag{119}$$

$$= \sup_{\lambda\geqslant 0,\eta\in\mathbb{R},\frac{-H(x)+\eta}{\lambda}\leqslant 1} \{-\mathbb{E}_{P_0}\Big[\max\Big\{-H(X)+\eta, -\lambda\Big\}\Big] - \lambda\rho + \eta\} \tag{120}$$

$$= \sup_{\lambda\geqslant 0,\eta\in\mathbb{R},\frac{-H(x)+\eta}{\lambda}\leqslant 2} \{-\mathbb{E}_{P_0}\Big[\max\Big\{-H(X)+\eta-\lambda, -\lambda\Big\}\Big] - \lambda\rho + \eta - \lambda\} \tag{121}$$

$$= \sup_{\lambda\geqslant 0,\eta\in\mathbb{R},\frac{-H(x)+\eta}{\lambda}\leqslant 2} \{-\mathbb{E}_{P_0}\Big[\max\Big\{-H(X)+\eta, 0\Big\}\Big] - \lambda\rho + \eta\} \tag{122}$$

$$= \sup_{\lambda\geqslant 0,\eta\in\mathbb{R},\frac{-H(x)+\eta}{\lambda}\leqslant 2} \{-\mathbb{E}_{P_0}\Big[-H(X)+\eta\Big]_+ - \lambda\rho + \eta\}. \tag{123}$$

Here Equation (121) is obtained by substituting η with $\eta-\lambda$. In order to optimize over λ, we need to choose the minimum λ satisfying the constraints. We require $\lambda \geqslant \frac{-H(x)+\eta}{2}$ which translates to $\lambda \geqslant \frac{-ESI(H(x))+\eta}{2}$ (as this constraint originates inside the expectation, points with zero mass, $\{t\in\mathbb{R}:\mathbb{P}\{Y<t\}=0\}$, will have no effect). Substituting this, we have

$$\inf_{P:\mathrm{TV}(P||P_0)\leqslant\rho} \mathbb{E}_P[H(X)] = \sup_{\eta\in\mathbb{R}}\{-\mathbb{E}_{P_0}\Big[-H(X)+\eta\Big]_+ - \frac{(-ESI(H(x))+\eta)_+}{2}\rho + \eta\}. \tag{124}$$

Denote the inner function in Equation (124), as

$$g_{\mathrm{TV}}(\eta, P_0) = -\mathbb{E}_{P_0}\Big[-H(X)+\eta\Big]_+ - \frac{(-ESI(H(x))+\eta)_+}{2}\rho + \eta. \tag{125}$$

Note that for $\eta\leqslant 0$, the first two terms in $g_{\mathrm{TV}}(\eta,P_0)$ will be 0 if $H(x)>0$ for all x. This implies

$$g_{\mathrm{TV}}(\eta, P_0) = \eta \leqslant 0 \quad \forall \quad \eta\leqslant 0. \tag{126}$$

Also, as $H(x)\leqslant\frac{1}{1-\gamma}$, we substitute $\eta=\frac{2+\rho}{\rho(1-\gamma)}$ in $g_{\mathrm{TV}}(\eta,P_0)$, and bound it as follows:

$$g_{\mathrm{TV}}\Big(\frac{(2+\rho)}{\rho(1-\gamma)}, P_0\Big) = -\mathbb{E}_{P_0}\Big[-H(X)+\frac{(2+\rho)}{\rho(1-\gamma)}\Big]_+ - \frac{(-ESI(H(x))+\frac{(2+\rho)}{\rho(1-\gamma)})_+}{2}\rho + \frac{(2+\rho)}{\rho(1-\gamma)} \tag{127}$$

$$= \mathbb{E}_{P_0}\Big[H(X)\Big] - \frac{(2+\rho)}{\rho(1-\gamma)} - \frac{(-ESI(H(x))+\frac{(2+\rho)}{\rho(1-\gamma)})_+}{2}\rho + \frac{(2+\rho)}{\rho(1-\gamma)} \tag{128}$$

$$= \mathbb{E}_{P_0}\Big[H(X)\Big] - \frac{(-ESI(H(x))+\frac{(2+\rho)}{\rho(1-\gamma)})_+}{2}\rho \tag{129}$$

$$= \mathbb{E}_{P_0}\Big[H(X)\Big] - \frac{(-ESI(H(x))+\frac{(2+\rho)}{\rho(1-\gamma)})}{2}\rho \tag{130}$$

$$= \mathbb{E}_{P_0}\Big[H(X)-\frac{1}{1-\gamma}\Big] + \frac{\rho ESI(H(x))}{2} - \frac{\rho}{2(1-\gamma)} \tag{131}$$

$$= \mathbb{E}_{P_0}\Big[H(X)-\frac{1}{1-\gamma}\Big] + \frac{\rho}{2}(ESI(H(x)) - \frac{1}{(1-\gamma)}) \tag{132}$$

$$\leqslant 0. \tag{133}$$

Here Equation (128), Equation (130) and Equation (133) are obtained from the fact that that $H(x)\leqslant\frac{1}{1-\gamma}$ $(-H(x)+\frac{(2+\rho)}{\rho(1-\gamma)}>0)$ and $ESI(H(x))\leqslant\frac{1}{1-\gamma}$ $(-ESI(H(x))+\frac{(2+\rho)}{\rho(1-\gamma)}>0)$. Along with the convexity of $g_{\mathrm{TV}}(\eta,P_0)$, Equation (126) and Equation (133) imply that the $\sup_{\eta\in\mathbb{R}}\{g_{\mathrm{TV}}(\eta,P_0)\}$ is attained in the η range $[0,\frac{(2+\rho)}{\rho(1-\gamma)}]$. □

Theorem 4. *(Sample Complexity under TV uncertainty set) Consider a robust MDP (see Section 2) with nominal transition dynamics f and uncertainty set defined as in Equation (2) w.r.t. TV distance. For π^* denoting the robust optimal policy w.r.t. nominal transition dynamics f and π^*_N denoting the robust optimal policy w.r.t. learned nominal transition dynamics $\hat{f}_N$ via Algorithm 1, and $\delta \in (0,1)$, $\epsilon \in (0, \frac{1}{1-\gamma})$,it holds that $\max_s |V^R_{\pi^*_N,f}(s) - V^R_{\pi^*,f}(s)| \leqslant \epsilon$ with probability at least $1-\delta$ for any $N \geqslant N_{\mathrm{TV}}$, where*

$$N_{\mathrm{TV}} = \mathcal{O}\Big(\frac{(2+\rho)^2\gamma^2}{\rho^2(1-\gamma)^4}\frac{\beta_n(\delta)^2 d^2 \gamma_{nd}}{\epsilon^2}\Big). \tag{134}$$

Proof. **Step (i):** As detailed in the proof outline of Section 4, in order to bound $V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)$, we begin by adding and subtracting $V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s)$ which is the robust value function w.r.t. the nominal transition dynamics $\hat{f}_n$ and its corresponding optimal policy $\hat{\pi}_n$. Then, we split the difference into two terms as follows:

$$V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\pi^*,f}(s) = \underbrace{V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s)}_{(i)} + \underbrace{V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)}_{(ii)}. \tag{135}$$

In order to not disturb the flow of the proof we bound (i) and (ii) separately Lemma 6 and Lemma 7 respectively. From Lemma 6, we obtain that

$$\begin{aligned}(i) &\leqslant \max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s)\Big| \\ &\leqslant \frac{\gamma}{1-\gamma}\max_s \Big|\inf_{\mathrm{TV}(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big] - \inf_{\mathrm{TV}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big]\Big|.\end{aligned} \tag{136}$$

And from Lemma 7, we obtain that

$$\begin{aligned}(ii) &\leqslant \max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)\Big| \\ &\leqslant \frac{\gamma}{1-\gamma}\max_s \Big|\inf_{\mathrm{TV}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\pi^*,f}(s')\Big] - \inf_{\mathrm{TV}(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\pi^*,f}(s')\Big]\Big|.\end{aligned} \tag{137}$$

Note that both these terms in Equations (136) and (137) are of the form mentioned in the **Step (i)** of Section 4.

Step (ii): Next, corresponding to **step (ii)** of the proof outline in Section 4, we use Lemma 19 to bound Equations (136) and (137). Denote $M := \frac{1}{1-\gamma} \geqslant \max_s V^{\mathrm{R}}_{\pi}(s)$ for convenience. Using Equation (136) and Lemma 20 (internally using Lemma 19), it holds that

$$\begin{aligned}(i) &\leqslant \max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s)\Big| \\ &\leqslant \frac{1}{1-\gamma}\max_s \Big|\gamma \inf_{\mathrm{TV}(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big] - \gamma \inf_{\mathrm{TV}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V^{\mathrm{R}}_{\hat{\pi}_n,f}(s')\Big]\Big| \\ &\leqslant \frac{\gamma}{1-\gamma}\max_{s,a}\left(\sup_{\eta\in[0,\frac{(2+\rho)}{\rho(1-\gamma)}]}\Big\{\Big|(\mathbb{E}_{P_f(s,a)}[(-V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s')+\eta)_+]) - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s')+\eta)_+])\Big|\Big\}\right) \qquad (138) \\ &\leqslant \frac{\gamma}{1-\gamma}\max_{V(\cdot)\in\mathcal{V}}\max_{s,a}\left(\sup_{\eta\in[0,\frac{(2+\rho)}{\rho(1-\gamma)}]}\Big\{\Big|(\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)_+]) - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')+\eta)_+])\Big|\Big\}\right). \qquad (139)\end{aligned}$$

We can bound (ii) similarly.

$$\begin{aligned}(ii) &\leqslant \max_s \Big|V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) - V^{\mathrm{R}}_{\pi^*,f}(s)\Big| \qquad (140) \\ &\leqslant \frac{\gamma}{1-\gamma}\max_{V(\cdot)\in\mathcal{V}}\max_{s,a}\left(\sup_{\eta\in[0,\frac{(2+\rho)}{\rho(1-\gamma)}]}\Big\{\Big|(\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)_+]) - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')+\eta)_+])\Big|\Big\}\right). \qquad (141)\end{aligned}$$

Step (iii): Next, we want to utilize the learning error bound (Equation (26)) that bounds the difference between the means of true nominal transition dynamics P_f and learned nominal transition dynamics $P_{\hat{f}_n}$ to bound Equations (139) and (141).

We begin by bounding the difference $\left|\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)_+] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')+\eta)_+]\right|$, by the difference in means of P_f and $P_{\hat{f}_n}$ in Lemma 21. Since Equation (139) has a max over all value functions, we introduce a covering number argument in Lemma 22 to reform it to a max over the functions in the $\zeta-$covering set. We then use Lemma 21 to obtain bounds in terms of maximum information gain Γ_{Nd} (Equation (9)) and ζ. Further details regarding the covering number argument are deferred to Lemma 22. Then, we apply the result of Lemma 22 with $\zeta = 1$ (defined in Lemma 22) on Equation (139). Then, it holds that

$$(i) \leqslant \max_s \left|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s)\right| = \mathcal{O}\left(\left(\tfrac{(2+\rho)\gamma}{\rho(1-\gamma)^2}\right)\left(\tfrac{\beta_n(\delta)\sqrt{2ed^2\gamma_{nd}}}{\sigma\sqrt{n}}\right)\right). \tag{142}$$

Note that β_n, which appears in Lemma 3, has a logarithmic dependence on n. Similarly, from Equation (141), and Lemmas 21 and 22, we obtain

$$(ii) \leqslant \max_s \left|V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) - V^{\mathrm{R}}_{\pi*,f}(s)\right| = \mathcal{O}\left(\left(\tfrac{(2+\rho)\gamma}{\rho(1-\gamma)^2}\right)\left(\tfrac{\beta_n(\delta)\sqrt{2ed^2\gamma_{nd}}}{\sigma\sqrt{n}}\right)\right). \tag{143}$$

Note that we want to bound $V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\pi*,f}(s) = (i) + (ii)$ over all $s \in \mathcal{S}$. Using $\max_s \left|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\pi*,f}(s)\right| \leqslant \max_s \left|V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) - V^{\mathrm{R}}_{\pi*,f}(s)\right| + \max_s \left|V^{\mathrm{R}}_{\hat{\pi}_n,\hat{f}_n}(s) - V^{\mathrm{R}}_{\pi^*_n,f}(s)\right|$ and substituting M by $1/(1-\gamma)$, we obtain from Equation (142) and Equation (143)

$$\max_s \left|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\pi*,f}(s)\right| = \mathcal{O}\left(\left(\tfrac{(2+\rho)\gamma}{\rho(1-\gamma)^2}\right)\left(\tfrac{\beta_n(\delta)\sqrt{2ed^2\gamma_{nd}}}{\sigma\sqrt{n}}\right)\right).$$

Finally, to ensure that $\max_s |V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\pi*,f}(s)| \leqslant \epsilon$, it suffices to have

$$\max_s \left|V^{\mathrm{R}}_{\hat{\pi}_n,f}(s) - V^{\mathrm{R}}_{\pi*,f}(s)\right| = \mathcal{O}\left(\left(\tfrac{(2+\rho)\gamma}{\rho(1-\gamma)^2}\right)\left(\tfrac{\beta_n(\delta)\sqrt{2ed^2\gamma_{nd}}}{\sigma\sqrt{n}}\right)\right) = \epsilon.$$

Moving $\sqrt{n}$ and ϵ to opposite sides and squaring both sides, we obtain

$$n = \mathcal{O}\left(\left(\frac{(2+\rho)^2\gamma^2}{\rho^2(1-\gamma)^4}\right)\left(\frac{\beta_n(\delta)^2 2ed^2\gamma_{nd}}{\sigma^2\epsilon^2}\right)\right).$$

□

Lemma 20. *(Simplification using Lemma 19 reformulation) Let V be a value function from $\mathcal{S} \to [0, 1/(1-\gamma)]$. Then, it holds that*

$$\max_s \Big|\inf_{\mathrm{TV}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V(s')\Big] - \inf_{\mathrm{TV}(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V(s')\Big]\Big| \leqslant \max_{s,a} \sup_{\eta\in[0,\frac{(2+\rho)}{\rho(1-\gamma)}]} \{|(\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)_+]) - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')+\eta)_+])|\}.$$

Proof. First note that,

$$\max_s \Big|\inf_{\mathrm{TV}(p||P_{\hat{f}_n}(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V(s')\Big] - \inf_{\mathrm{TV}(p||P_f(s,\hat{\pi}_n(s)))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V(s')\Big]\Big| \leqslant \max_{s,a} \Big|\inf_{\mathrm{TV}(p||P_{\hat{f}_n}(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V(s')\Big] - \inf_{\mathrm{TV}(p||P_f(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V(s')\Big]\Big| \tag{144}$$

Using Lemma 19 and focusing to bound right side of Equation (144) for one particular (s,a) state action pair, we obtain

$$\Big|\inf_{\mathrm{TV}(p||P_{\hat{f}_n}(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V(s')\Big] - \inf_{\mathrm{TV}(p||P_f(s,a))\leqslant\rho} \mathbb{E}_{s'\sim p}\Big[V(s')\Big]\Big|$$

$$= \Big| \sup_{\eta\in[0,\frac{(2+\rho)}{\rho(1-\gamma)}]} \{-\mathbb{E}_{P_f(s,a)}\Big[-V(s')+\eta\Big]_+ - \frac{(-ESI_{P_f(s,a)}(V(s'))+\eta)_+}{2}\rho+\eta\} - \sup_{\eta\in[0,\frac{(2+\rho)}{\rho(1-\gamma)}]} \{-\mathbb{E}_{P_{\hat{f}_n}(s,a)}\Big[-V(s')+\eta\Big]_+ - \frac{(-ESI_{P_{\hat{f}_n}(s,a)}(V(s'))+\eta)_+}{2}\rho+\eta\}\Big| \tag{145}$$

$$\leqslant \sup_{\eta\in[0,\frac{(2+\rho)}{\rho(1-\gamma)}]} \{|(\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)_+]) - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')+\eta)_+])|\}. \tag{146}$$

Here, Equation (146) is obtained using $ESI_{P_f(s,a)}(V(s')) = ESI_{P_{\hat{f}_n}(s,a)}(V(s'))$ as shown in proof of Lemma 9 (Case-1).

□

Lemma 21. *(Bound by difference between estimated model $\hat{f}_n$ and true f) Let V be a value function from $\mathcal{S}\to[0,1/(1-\gamma)]$. Then, it holds that*

$$|\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)_+] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')+\eta)_+]| \leqslant \Big(\frac{(2+\rho)}{\rho(1-\gamma)}\Big)\sigma^{-1}\|f(s,a)-\hat{f}_n(s,a)\|, \tag{147}$$

where $P_{\hat{f}_n}(s,a) = \mathcal{N}(\hat{f}_n(s,a),\sigma^2 I)$ and $P_f(s,a) = \mathcal{N}(f(s,a),\sigma^2 I)$ and $\eta\in[0,\frac{(2+\rho)}{\rho(1-\gamma)}]$.

Proof.

$$\begin{aligned}
&\Big|\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)_+] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')+\eta)_+]\Big| \\
&= \Big|\int_{\mathbb{R}^d} \frac{1}{\sqrt{(2\pi\sigma^2)^d}}(-V(s')+\eta)_+(e^{-\frac{\|x-f(s,a)\|^2}{2\sigma^2}} - e^{-\frac{\|x-\hat{f}_n(s,a)\|^2}{2\sigma^2}})\Big| \\
&\leqslant \int_{\mathbb{R}^d} \frac{1}{\sqrt{(2\pi\sigma^2)^d}}(-V(s')+\eta)_+\Big|e^{-\frac{\|x-f(s,a)\|^2}{2\sigma^2}} - e^{-\frac{\|x-\hat{f}_n(s,a)\|^2}{2\sigma^2}}\Big| \\
&\overset{(i)}{\leqslant} \frac{(2+\rho)}{\rho(1-\gamma)}\int_{\mathbb{R}^d} \frac{1}{\sqrt{(2\pi\sigma^2)^d}}\Big|e^{-\frac{\|x-f(s,a)\|^2}{2\sigma^2}} - e^{-\frac{\|x-\hat{f}_n(s,a)\|^2}{2\sigma^2}}\Big| \\
&\overset{(ii)}{\leqslant} 2\frac{(2+\rho)}{\rho(1-\gamma)}\cdot \mathrm{TV}(P_{\hat{f}_n}(s,a), P_f(s,a)) \\
&\overset{(iii)}{\leqslant} 2\frac{(2+\rho)}{\rho(1-\gamma)}\sqrt{\mathrm{KL}(P_{\hat{f}_n}(s,a), P_f(s,a))/2} \\
&\overset{(iv)}{\leqslant} 2\frac{(2+\rho)}{\rho(1-\gamma)}\sqrt{\|f(s,a)-\hat{f}_n(s,a)\|^2/4\sigma^2} \\
&\leqslant \frac{(2+\rho)}{\rho(1-\gamma)}\|f(s,a)-\hat{f}_n(s,a)\|/\sigma,
\end{aligned}$$

where (i) follows from $(-V(s')+\eta)_+^2 \leqslant \frac{(2+\rho)}{\rho(1-\gamma)}$ as $\eta\leqslant\frac{(2+\rho)}{\rho(1-\gamma)}$, (ii) follows from the definition of Total Variation (TV) distance between any two multivariate Gaussians, (iii) uses the Pinsker's inequality, and (iv) uses the formula for KL-divergence between multivariate Gaussian distributions.

□

Lemma 22. *($\zeta-$cover construction) For $\mathcal{V}$ denoting the set of value functions from $\mathcal{S}\to[0,1/(1-\gamma)]$, with probability at least $1-\delta$ it holds that*

$$\max_{V\in\mathcal{V}}\max_{s,a}\sup_{\eta\in[0,\frac{(2+\rho)}{\rho(1-\gamma)}]}\{|\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)_+] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')+\eta)_+]|\} \leqslant \mathcal{O}\left(\Big(\frac{(2+\rho)}{\rho(1-\gamma)}\Big)\Big(\frac{\beta_n(\delta)\sqrt{2ed^2\gamma_{nd}}}{\sigma\sqrt{n}}\Big)\right). \tag{148}$$

Proof. Let $\mathcal{N}_{\mathcal{V}}(\zeta)$ be the $\zeta-$ cover of the set $\mathcal{V}$. By definition, there exists $V' \in \mathcal{N}_{\mathcal{V}}(\zeta)$ such that $\|V' - V\| \leqslant \zeta$ for every $V \in \mathcal{V}$.

$$\begin{aligned}
&|\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)_+] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')+\eta)_+]| \\
&\leqslant |\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)_+] - \mathbb{E}_{P_f(s,a)}[(-V'(s')+\eta)_+]| \\
&\quad + |\mathbb{E}_{P_f(s,a)}[(-V'(s')+\eta)_+] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V'(s')+\eta)_+]| \\
&\quad + |\mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V'(s')+\eta)_+] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')+\eta)_+]|. \qquad (149)\\
&\overset{(i)}{\leqslant} 2\|V'-V\| + |\mathbb{E}_{P_f(s,a)}[(-V'(s')+\eta)_+] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V'(s')+\eta)_+]|, \qquad (150)
\end{aligned}$$

where (i) follows from Lemma 23. Using Equation (150), we bound uniformly over all $V \in \mathcal{V}$. Using Equation (150) we bound uniformly over all $V \in \mathcal{V}$,

$$\begin{aligned}
&\max_{V\in\mathcal{V}} \max_{s,a} \sup_{\eta\in[0,\frac{(2+\rho)}{\rho(1-\gamma)}]} \{|\mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)_+] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V(s')+\eta)_+]|\} \qquad (151)\\
&\leqslant \max_{V'\in\mathcal{N}_{\mathcal{V}}(\zeta)} \max_{s,a} \sup_{\eta\in[0,\frac{(2+\rho)}{\rho(1-\gamma)}]} \left\{ \left| 2\|V'-V\| + |\mathbb{E}_{P_f(s,a)}[(-V'(s')+\eta)_+] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V'(s')+\eta)_+]| \right| \right\} \\
&\overset{(ii)}{\leqslant} \max_{V'\in\mathcal{N}_{\mathcal{V}}(\zeta)} \max_{s,a} \sup_{\eta\in[0,\frac{(2+\rho)}{\rho(1-\gamma)}]} \left\{ \left| \mathbb{E}_{P_f(s,a)}[(-V'(s')+\eta)_+] - \mathbb{E}_{P_{\hat{f}_n}(s,a)}[(-V'(s')+\eta)_+] \right| \right\} + 2\zeta \\
&\overset{(iii)}{\leqslant} \max_{V'\in\mathcal{N}_{\mathcal{V}}(\zeta)} \max_{s,a} \sup_{\eta\in[0,\frac{(2+\rho)}{\rho(1-\gamma)}]} \left\{ \left(\tfrac{(2+\rho)}{\rho(1-\gamma)} \right) \sigma^{-1} \|f(s,a) - \hat{f}_n(s,a)\| \right\} + 2\zeta \\
&\overset{(iv)}{\leqslant} \mathcal{O}\left(\left(\tfrac{(2+\rho)}{\rho(1-\gamma)} \right) \left(\tfrac{\beta_n(\delta)\sqrt{2ed^2\gamma_{nd}}}{\sigma\sqrt{n}} \right) \right) + 2\zeta \qquad (152)\\
&\overset{(v)}{\leqslant} \mathcal{O}\left(\left(\tfrac{(2+\rho)}{\rho(1-\gamma)} \right) \left(\tfrac{\beta_n(\delta)\sqrt{2ed^2\gamma_{nd}}}{\sigma\sqrt{n}} \right) \right), \qquad (153)
\end{aligned}$$

where (ii) follows from $\|V' - V\| \leqslant \zeta$, (iii) follows from Lemma 21, (iv) follows from Equation (26), and (v) follows from substituing $\zeta = 1$ (or any constant).

□

Lemma 23. *For any two value functions $V, V' : \mathcal{S} \to [0, \frac{1}{1-\gamma}]$, it holds that*

$$|\mathbb{E}_{P_f(s,a)}[(-V'(s')+\eta)_+] - \mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)_+]| \leqslant \|V'-V\|. \qquad (154)$$

Proof. Noting that both the distributions are w.r.t. the same distribution $P_f(s,a)$ we have,

$$\begin{aligned}
&\mathbb{E}_{P_f(s,a)}[(-V'(s')+\eta)_+] - \mathbb{E}_{P_f(s,a)}[(-V(s')+\eta)_+] \\
&\leqslant \int_{s'\sim P_f(s,a)} \Big(\mathbb{1}(V'(s') < \eta)(-V'(s')+\eta) - \mathbb{1}(V(s') < \eta)(-V(s')+\eta) \Big) p_{P_f(s,a)}(s') ds'. \qquad (155)
\end{aligned}$$

Adding and subtracting $\mathbb{1}(V(s') < \eta)(-V'(s')+\eta)$ to Equation (155), we obtain 2 terms,

$$i = \int_{s'\sim P_f(s,a)} \Big(\mathbb{1}(V'(s') < \eta) - \mathbb{1}(V(s') < \eta) \Big) (-V'(s')+\eta) p_{P_f(s,a)}(s') ds' \qquad (156)$$

$$ii = \int_{s'\sim P_f(s,a)} \mathbb{1}(V(s') < \eta) \Big((-V'(s')+\eta) - (-V(s')+\eta) \Big) p_{P_f(s,a)}(s') ds'. \qquad (157)$$

Bounding i first,

$$i = \int_{s'\sim P_f(s,a)} \Big(\mathbb{1}(V'(s') < \eta) - \mathbb{1}(V(s') < \eta) \Big) (-V'(s')+\eta) p_{P_f(s,a)}(s') ds' \qquad (158)$$

$$= \int_{s' \sim P_f(s,a)} \Big(\mathbb{1}(V'(s') < \eta \leqslant V(s')) \Big)(-V'(s') + \eta) p_{P_f(s,a)}(s') ds' \tag{159}$$
$$- \int_{s' \sim P_f(s,a)} \Big(\mathbb{1}(V(s') < \eta < V'(s')) \Big)(-V'(s') + \eta) p_{P_f(s,a)}(s') ds'$$
$$\leqslant \int_{s' \sim P_f(s,a)} \Big(\mathbb{1}(V'(s') < \eta \leqslant V(s')) \Big)(-V'(s') + V(s')) p_{P_f(s,a)}(s') ds' \tag{160}$$
$$- \int_{s' \sim P_f(s,a)} \Big(\mathbb{1}(V(s') < \eta < V'(s')) \Big)(-V'(s') + V(s')) p_{P_f(s,a)}(s') ds'$$
$$\leqslant \int_{s' \sim P_f(s,a)} \Big(\mathbb{1}(V'(s') < \eta \leqslant V(s')) \Big)(-V'(s') + V(s')) p_{P_f(s,a)}(s') ds' \tag{161}$$
$$+ \int_{s' \sim P_f(s,a)} \Big(\mathbb{1}(V(s') < \eta < V'(s')) \Big)(V'(s') - V(s')) p_{P_f(s,a)}(s') ds'$$
$$\leqslant \|V' - V\|. \tag{162}$$

Similarly bounding ii,

$$ii = \int_{s' \sim P_f(s,a)} \mathbb{1}(V(s') < \eta) \Big((-V'(s') + \eta) - (-V(s') + \eta) \Big) p_{P_f(s,a)}(s') ds' \tag{163}$$
$$= \int\limits_{s' \sim P_f(s,a)} \mathbb{1}(V(s') < \eta) \Big(- V'(s') + V(s') \Big) p_{P_f(s,a)}(s') ds' \tag{164}$$
$$\leqslant \int\limits_{s' \sim P_f(s,a)} \mathbb{1}(V(s') < \eta) \Big| - V'(s') + V(s') \Big| p_{P_f(s,a)}(s') ds' \tag{165}$$
$$\leqslant \|V' - V\|. \tag{166}$$

Using Equations (162) and (166) we get the desired result. □

D Additional Experiments and Details

In this section, we report additional experiments and discuss further details of our experimental setup. All experiments were run with GPU clusters: 10xNVidia 32Gb Tesla V100 with Intel(R) processors (2 cores, 2.50 GHz) and 256Gb RAM. For all the experiments, we use the environment implementations of Mehta et al. (2021) as done in https://github.com/fusion-ml/trajectory-information-rl/tree/main. Also, to learn the environment transition model, we use the same corresponding GP hyperparameters proposed by Mehta et al. (2021). For the offline RFQI/FQI algorithms we follow the implementation of Panaganti et al. (2022); Chen and Jiang (2019) in https://github.com/zaiyan-x/RFQI. We use the same default hyperparameters as used in their code except for training steps, batch size and robustness radius ρ (for RFQI) which we tune depending on the environment as outlined next. For SAC in Pendulum experiments, we use the implementation and hyperparameters of https://github.com/DLR-RM/rl-baselines3-zoo. Whereas, for SAC in Reacher experiments, we use the implementation and hyperparameters of https://github.com/fusion-ml/bac-baselines, https://github.com/IanChar/rlkit2 (as done in (Mehta et al., 2021)).

Pendulum: In Pendulum experiments, we construct the learned model using 60 samples from the true environment. Then, we train a SAC policy on such a model for $2 * 10^4$ steps and use it (with the probability of choosing a random action being 0.3 or 0.5) to generate 10^6 offline data (these are used both for MVR+RFQI and MVR+FQI). For training steps and batch size we consider the following combinations: $\{'2000 - 100', '5000 - 100', '10000 - 100', '20000 - 100', '35000 - 100', '50000 - 100', '5000 - 500', '5000 - 1000'\}$. We combine all these combinations with the following values of ρ – $\{0.1, 0.2, 0.3, 0.5, 0.6, 0.7, 0.8, 0.9\}$. For each algorithm, we pick the best-performing combination in terms of average reward over 20 episodes for all (or most) perturbation values. We do this separately for length perturbations and action perturbations. In the length perturbation, the pendulum's length is changed from its nominal value to a new value depending on the perturbation percentage. In the action perturbation, a random action is chosen instead of the action chosen by the policy with various probabilities ranging from $[0, 1]$. We detail the optimal hyperparameters we realized for each algorithm in Table 2 for the

length and action perturbation, respectively. Moreover, we plot the average performance (over 20 episodes) of the different baselines w.r.t. length and action perturbations in Figure 3. We notice that in the case of length perturbation, the robust algorithms (RFQI and MVR+RFQI) outperform the corresponding non-robust baselines. In the case of action perturbations, we observe all algorithms except for SAC achieve similar performance.

	Training Steps	Batch-Size	ρ	Random Action Probability (Dataset)
MVR+RFQI	5000	100	0.3	0.5
MVR+FQI	2000	100	-	0.5
RFQI	2000	100	0.9	0.5
FQI	5000	500	-	0.5

	Training Steps	Batch-Size	ρ	Random Action Probability (Dataset)
MVR+RFQI	20000	100	0.5	0.3
MVR+FQI	50000	100	-	0.3
RFQI	50000	100	0.1	0.5
FQI	5000	500	-	0.5

Table 2: Hyperparameters for Pendulum - length perturbation (top) and action perturbation (bottom).

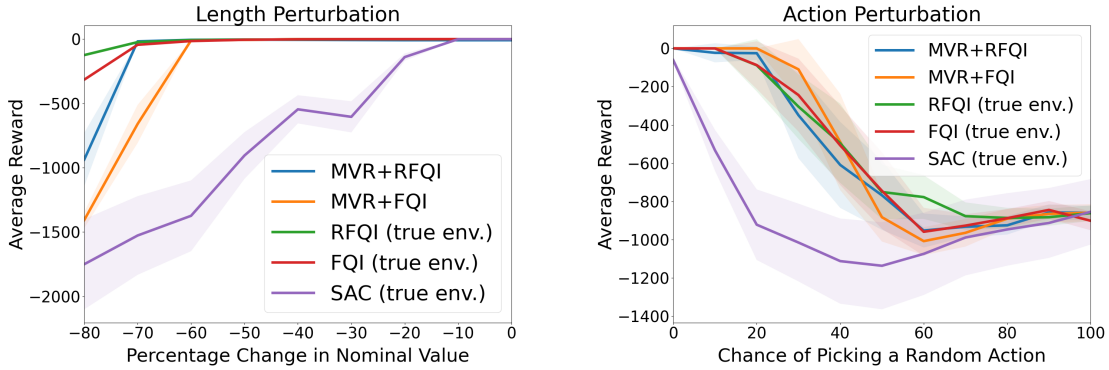


Figure 3: Pendulum experiments.

Cartpole: In Cartpole experiments, we construct the learned model using 150 samples from the true environment. Then, we run MPC on such a model following the implementation and hyperparameters of (Mehta et al., 2021; Pinneri et al., 2020) requiring 2250 samples to calculate the optimal action at each step and use it (with the probability of choosing a random action being 0.3) to generate 10^6 offline data for MVR+RFQI and MVR+FQI. For training steps and batch size, we test the following combinations: $\{'2000-100','5000-100','10000-100','20000-100','35000-100','50000-100','5000-500','5000-1000'\}$, and consider radii ρ in $\{0.1, 0.2, 0.3, 0.5, 0.6, 0.7, 0.8, 0.9\}$. We consider perturbations of the force magnitude and the gravity, whereby the actuation force/gravity is changed from its nominal value to a new value depending on the perturbation percentage. We report the best-performing (average over 20 episodes) hyperparameters for each algorithm in Table 3. Such parameters were observed to be a good choice for both perturbation types. Finally, we plot the average performance (over 20 episodes) of the different baselines w.r.t. force magnitude and gravity perturbations in Figure 4. We notice that in both perturbations, the robust algorithms (RFQI and MVR+RFQI) outperform the corresponding non-robust baselines.

Reacher: In Reacher experiments, we construct the learned model using 2000 samples from the true environment. Then, we train a SAC policy on such a model for 10^6 steps and use it (with the probability of choosing a random action being 0.3) to generate 10^6 offline data for MVR+RFQI and MVR+FQI. For training steps and batch size, we consider the following combinations: $\{'10000-500','20000-500','40000-500','80000-500','160000-1000'\}$, while we consider radii ρ in $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. We consider perturbations of the joint stiffness subject to different equilibrium positions, the latter represented by the 'Springref' parameter which we take to be 50 or 100. In both perturbation types, the joint stiffness is changed from its nominal value of 0 to a new value depending on

	Training Steps	Batch-Size	ρ	Random Action Probability (Dataset)
MVR+RFQI	5000	500	0.5	0.3
MVR+FQI	50000	100	-	0.3
RFQI	5000	100	0.3	0.3
FQI	10000	100	-	0.3

Table 3: Hyperparameters for Cartpole.

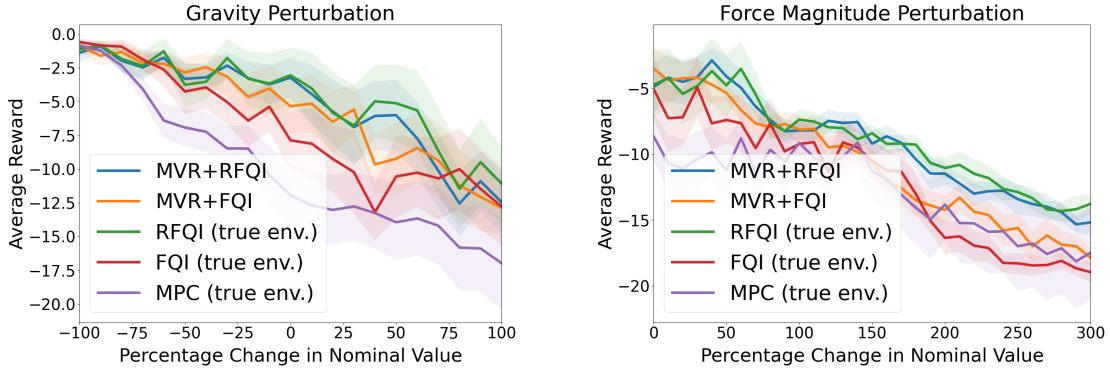


Figure 4: Cartpole experiments.

the perturbation magnitude. Best-performing hyperparameters' configurations are reported in Table 4. We plot the average performance (over 20 episodes) of the different baselines in Figure 5. Similar to the other environments, we observe the robust algorithms (RFQI and MVR+RFQI) outperform the corresponding non-robust baselines.

	Training Steps	Batch-Size	ρ	Random Action Probability (Dataset)
MVR+RFQI	10000	500	0.5	0.3
MVR+FQI	20000	500	-	0.3
RFQI	40000	500	0.1	0.3
FQI	20000	500	-	0.3

Table 4: Hyperparameters for Reacher.

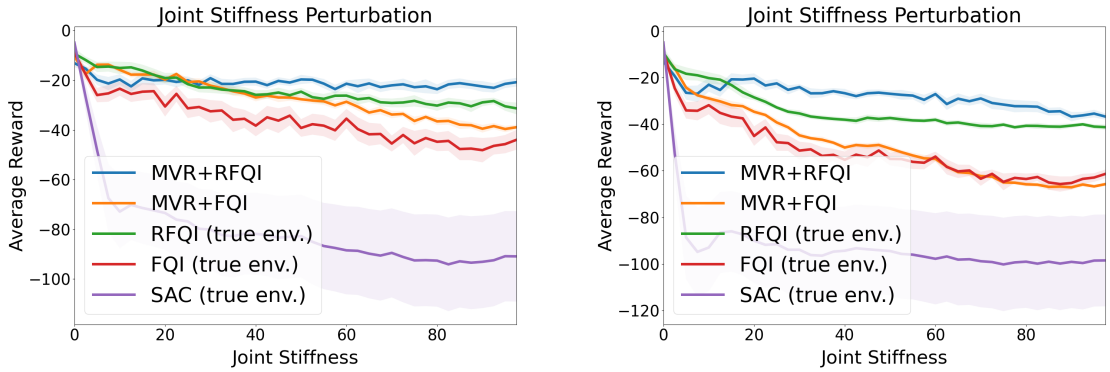


Figure 5: Reacher experiments with 'Springref' parameter set to 50 (left) or 100 (right).