Sample Complexity for Obtaining Sub-optimality and Violation bound for Distributionally Robust Constrained MDP

Arnob Ghosh
Department of Electrical and Computer Engineering
New Jersey Institute of Technology
arnob.ghosh@njit.edu

Abstract

We consider the problem of learning a safe policy that will maximize the cumulative reward while satisfying a constraint even when there is a mismatch between the testing and training environment. In particular, we consider the *robust* constrained Markov decision problem (CMDP) where an agent needs to maximize the reward and satisfy the constraint against the worst possible stochastic model under the *unknown* uncertainty set. Such a problem poses significant additional challenges compared to the non-robust CMDP problem and the unconstrained robust MDP problem. We seek to characterize the number of samples required to bound both the sub-optimality gap and the violations by at most ϵ. We observe that the primal-dual-based approaches that achieves sample complexity bounds for non-robust CMDP cannot achieve the same in the robust CMDP case as the strong duality does not exist even when Slater's condition is satisfied. Nevertheless, we propose a robust safe value learning algorithm by considering an approach where *rectified* penalty for the constraint violation is added with the objective. We consider that the algorithm has access to a generative model of the *nominal* (training) environment around which the uncertainty set is defined. We show that our proposed algorithm can achieve a policy with ϵ suboptimality gap and violation bound after $\tilde{\mathcal{O}}(H^5|S||A|/\epsilon^2)$ samples; where $|S|$ is the cardinality of the state-space, $|A|$ is the cardinality of the action space, and H is the length of the episode for uncertainty sets specified by various popular distance metrics. *This is the first result that achieves a sample complexity bound for robust CMDP problems.*

1 Introduction

In many practical applications of online reinforcement learning (RL) (e.g., safety, resource constraints), there exist additional constraints on the learned policy in the sense that it also needs to ensure that the expected total utility (cost, resp.) exceeds a given threshold (is below a threshold, resp.). Such problems are formulated as constrained Markov Decision Processes (CMDPs) (Altman, 1999; Efroni et al., 2020). Algorithms are proposed with provable performance guarantee to solve for CMDP using both simulator or with online interaction (Vaswani et al., 2022; Ding et al., 2020; 2021).

However, a feasible policy for CMDP found by training using a simulator may violate the constraints when employed in the real environment because of the mismatch between the models. Such mismatch may exist because of the non-stationarity, sim-to-real gap, or even because of the adversarial attacks. Standard RL algorithms even fail to adapt to the model-mismatch in the unconstrained setting Sünderhauf et al. (2018); Tobin et al. (2017). *Thus, it is important to find a feasible (nearly) and yet close to the optimal policy even when there is a model mismatch using only simulator data.* Furthermore, we want to provide provable performance guarantee for the achieved policy. In

particular, when the uncertainty set on the models are unknown, and one is learning, we want to understand how many samples are required to attain such a policy.

While there are works that obtain sample complexity bound for unconstrained robust MDP, *it still remains open for robust CMDP setup*. Finding sample complexity bound for robust CMDP is inherently more challenging compared to the unconstrained case as one not only needs to find a policy that maximizes the worst-case, one also needs to ensure that the policy satisfies the constraint even when there is a mismatch in the model. Additional challenge arises as one even does not know the uncertainty set, rather it is learning. Recently, Sun et al. (2024) developed a policy-optimization based approach to improve the robust policy while ensuring feasibility. Wang et al. (2022) proposed primal-dual algorithm to obtain *local* solution. However, sample complexity bound have not been considered there. While primal-dual based and LP-based approaches are proposed for obtaining near-optimal sample complexity bound for non-robust CMDP, those approaches rely on the convexity of the state-action occupancy measure. However, Wang et al. (2022) shows that for the robust CMDP state-action occupancy measures are not convex. Hence, it is unclear on whether it is possible to achieve sample complexity bound for robust CMDP problem using primal-dual approach or traditional LP-based approach Efroni et al. (2020). We seek to answer the following open question–

Can we design an algorithm that achieves provable sample complexity result for robust CMDP?

Our Contributions: We consider an episodic robust CMDP setup where we seek to solve the following–

$$\max_{\pi} \min_{P\in\mathcal{P}} V_r^{\pi,P}(x) \quad \text{subject to } \min_{P\in\mathcal{P}} V_g^{\pi,P}(x) \geq b \tag{1}$$

where the objective is to maximize the worst case cumulative reward ($\min_P V_r^{\pi,P}(x)$) subject to the constraint that the worst case cumulative utility ($\min_P V_g^{P}(x)$) is above a certain threshold. We consider that the uncertainty set $\mathcal{P}$ of the transition probabilities P is unknown and is centered around the nominal model P^0. Similar to the unconstrained setup Panaganti et al. (2022); Xu et al. (2023); Gheshlaghi Azar et al. (2013); Kalathil et al. (2021), we consider that the agent *only* has access to the samples from this nominal model. We need to find a policy $\hat{\pi}$ after N_{tot} number of samples with the provable performance guarantee. In particular, we are interested in the following–

Definition 1. *We seek to obtain policy $\hat{\pi}$ such that after N_{tot} number of samples with high probability*

$$\text{Sub}-\text{Opt}(\hat{\pi}) := \min_P V_r^{\pi^*,P} - \min_P V_r^{\hat{\pi},P} \leq \epsilon \quad \text{Violation}(\hat{\pi}) := \min_P V_g^{\hat{\pi},P} \geq b - \epsilon \tag{2}$$

where π^* is the optimal policy for (1). Hence, we want to find the policy $\hat{\pi}$ such that it will be sub-optimal by at most ϵ amount and will violate by only ϵ amount. Note that compared to the unconstrained case, here, one also needs to bound violation.

Our main contributions are the followings:

1. We show that we can achieve $\text{Sub}-\text{Opt}(\hat{\pi}) \leq \epsilon$, and $\text{Violation}(\hat{\pi}) \leq \epsilon$ using $\tilde{\mathcal{O}}(H^5|S||A|/\epsilon^2)$ samples where H is the time-steps in an episode, $|S|$ is the cardinality of the state-space, and $|A|$ is the cardinality of the action-space. This is the *first finite sample complexity* result for robust CMDP, and yet matches the sample complexity guarantee for the unconstrained robust MDP. We completely characterize the sample complexity results for various uncertainty sets.

2. The traditional primal-dual-based approaches may be unable to achieve sample complexity results for the robust CMDP case unlike the non-robust case. This is because the strong duality result does not hold for the robust CMDP Wang et al. (2022). Further, while, the robust Bellman equation does hold for individual robust value function, it does not hold for the combined value functions since the worst case model may be different for the reward and the utility value function. Hence, one cannot apply the robust value iteration approach for the Lagrangian to solve for policy unlike the non-robust case. To address this, we propose a *rectified* penalty to consider the following

objective $\min_P V_r^{\pi,P}(x) - \lambda(b - \min_P V_g^{\pi,P}(x))_+$. We then consider an oracle that returns π solving the above for suitable choice of λ. It turns out that finding a policy that maximizes the above objective is key in obtaining the sample complexity bound for robust CMDP.

3. Our result does not depend on the strong duality unlike the approaches for the non-robust CMDP. Thus, our analysis techniques are also novel and hence can be of independent interest.

1.1 Related Literature

CMDP: Algorithms to compute policies for CMDP with provable performance guarantees are well-studied. For example, primal-dual based approaches with provable performance guarantee have been proposed Vaswani et al. (2022); Ghosh et al. (2022); Wei et al. (2022); Ding et al. (2020; 2021; 2023); Ghosh et al. (2024). The approaches used the strong duality result to show that the primal-dual approach converges to the optimal solution. Another popular methods including LP-based approaches can be classified as primal methods Xu et al. (2021); Efroni et al. (2020); Bura et al. (2022); Liu et al. (2021). The key idea idea was to use the convexity of the state-action occupancy measure. However, none of the works considered robust CMDP problem. Hence, the policy obtained by those approaches might not be feasible when there is a model mismatch. Further, as we described the robust state-action occupancy measure may not be convex as the worst-case transition kernel depends on the policy itself rendering the analysis inapplicable to obtain sample complexity bound.

Robust MDP: Robust MDP has been studied with *known* uncertainty set Iyengar (2005); Nilim & Ghaoui (2003); Wang & Zou (2021) and with unknown uncertainty set Panaganti et al. (2022); Xu et al. (2023); Yang & Wang (2019); Wang & Zou (2022); Ma et al. (2022); Zhou et al. (2021); Panaganti et al. (2022). Compared with the robust unconstrained MDP, robust MCDP is more challenging as one needs to ensure that the policy is feasible for every transition kernel under the uncertainty set. Directly applying the policy designed for unconstrained robust MDP will not satisfy the constraint for robust CMDP.

Robust CMDP: Unlike the non-robust CMDP, there are limited work on robust CMDP. Primal-dual based approach has been proposed in Russel et al. (2020); Mankowitz et al. (2020); Wang et al. (2022). However, Wang et al. (2022) showed that robust CMDP may not have strong duality guarantee even when a strictly feasible policy exists (aka Slater's condition) unlike the non-robust CMDP. Naturally, those approaches are unable to provide theoretical sample complexity bound. Recently, Sun et al. (2024) proposed an approach to improve the policy without violating the constraint, however, theoretical sample complexity bound has not been provided there.

2 Problem Formulation

Constrained Markov Decision Problem: An episodic constrained Markov Decision Process (CMDP) is characterized by the tuple $\{S, A, R, G, P, H\}$ where S is the state-space, A is the action-space; $R = \{r_h\}_{h=1}^H$ and $G = \{g_h\}_{h=1}^H$ are respectively the collection of rewards and utility for state-action pair (s, a) at step $h \in [H]$. H is the length of the episode. $P = \{P_h\}_{h=1}^H$ denotes the transition probability $P_{h,s,a}(s') = P_h(s'|s, a)$ at step h. Without loss of generality, we assume that r_h, and g_h are deterministic, and $|r_h(x, a)| \leq 1$, and $|g_h(x, a)| \leq 1$. We denote the policy $\pi_h(\cdot|x) \in \Delta(A)$ as the policy at time-step h. In a CMDP Efroni et al. (2020); Ghosh et al. (2022); Ding et al. (2021); Wei et al. (2022) setup one seeks to solve the following optimization problem.

$$\max_{\pi} V_r^{\pi,P}(x) \quad \text{subject to } V_g^{\pi,P}(x) \geq b \tag{3}$$

where $V_r^{\pi,P}(x) = \mathbb{E}_{\pi,P}[\sum_{h=1}^H r_h(x_h, a_h)|x_1 = x]$ and $V_g^{\pi}(x) = \mathbb{E}_{\pi,P}[\sum_{h=1}^H g_h(x_h, a_h)|x_1 = x]$ are the expected cumulative reward and the expected cumulative cost respectively following the policy π. We also denote $V_{j,h}^{\pi,P}(x) = \mathbb{E}_{\pi,P}[\sum_{i=h}^H j_i(x_i, a_i)|x_h = x]$ for $j = r, g$. Hence, $V_{j,1}^{\pi,P}(x) = V_j^{\pi,P}(x)$. The optimization problem in (3) denotes that we want to maximize the cumulative reward subject to the constraint that expected cumulative utility is above a certain threshold.

Example 1. *Consider the setup where the agent wants to maximize the reward while being at the safe state. In this case, the utility is $g_h(x) = 1$ if x is safe and 0 otherwise. This problem can be cast as a CMDP where $b = H$.*

Robust CMDP: We often use a simulator to train our policy before implementing in the real-life. However, the simulator setup and the real-life environment are often different, hence, we need a robust policy so that the policy can perform reasonably well in the real-life setup. In particular, we seek to solve the robust CMDP problem described in (1). $\rho > 0$, and is known.

In (1), $\mathcal{P}$ denotes the set of all transition probabilities. In particular, different transition probability defines different set of randomness inherent in the true environment. The problem in (1) defines that we seek to maximize the worst case expected cumulative reward subject to the constraint that the worst case cumulative utility is above the threshold b. *The objective of the robust CMDP formulation is that constraints are satisfied even if there are mismatch between training and evaluation the constraint is satisfied while maximizing the reward among the worst of all the transition probability model.* Such robustness guarantee is important for implementing RL algorithms in practice. Consider the example we described above, there, the solution in (1) ensures that the policy will still be safe even if there is a mismatch.

Note that our analysis and approach can be easily applicable to the setting where $\max_P V_g^{\pi,P} \leq b$ as well where g_h denotes the cost instead of utility at time-step h, and we are interested in the constraint such that the worst-case cost is below a certain threshold b. Further, we can easily extend our analysis for multiple constraints. *For notational simplicity, we interchangably denote $V_{j,h}^{\pi}(x) = \min_{P\in\mathcal{P}} V_{j,h}^{\pi,P}(x)$ for $j = r, g$, and all h.*

Uncertainty Set on models: Similar to the one considered in the unconstrained episodic MDP setup Xu et al. (2023), we consider a set of transition probability models within a ball centered around the nominal model $P^0_{h,s,a}$ $\forall (h,s,a) \in [H] \times S \times A$. We consider the uncertainty set $\mathcal{P} = \bigotimes_{(h,s,a)\in[H]\times S\times A} \mathcal{P}_{h,s,a}$ such that

$$\mathcal{P}_{h,s,a} = \{P \in \Delta(S) : D(P, P^0_{h,s,a}) \leq \rho\} \tag{4}$$

where D is the distance metric between two probability measures, and ρ is the radius of the uncertainty set. This uncertainty set satisfies the (s,a)-rectangularity assumption Iyengar (2005); Nilim & Ghaoui (2003). Our analysis can be extended trivially to s-rectangularity assumption as well Yang & Wang (2019). Without rectangualarity assumption, even for unconstrained robust MDP, obtaining policy is NP-hard problem Wiesemann et al. (2013). We do not assume that that we know the nominal model P^0, and thus we do not know the uncertainty set of transition kernels. We consider various uncertainty sets– i) Total variation uncertainty sets, ii) χ-squared uncertainty set, iii) KL-divergence uncertainty set, and iv) Wasserstein uncertainty set. Here, we describe the total variation uncertainty set, the rest are in Appendix D.

Total Variation uncertainty set: Let $\mathcal{P}^{TV} = \bigotimes_{(h,s,a)\in[H]\times S\times A} \mathcal{P}^{TV}_{h,s,a}$ be the uncertainty set defined in (4) with total variation distance Xu et al. (2023)

$$D_{TV}(P, P^0_{h,s,a}) = (1/2)||P - P^0_{h,s,a}||_1 \tag{5}$$

Generative Model: We do not know the uncertainty set, rather, we assume that we have access to a generative model or a simulator where the agent submits a query $(h,s,a) \in [H] \times S \times A$, and receives $s_{h+1} \sim P^0_{h,s,a}(\cdot)$, $r_h(s,a)$, and $g_h(s,a)$. Accessing the generative model or simulator is a common assumption even for unconstrained robust MDP Xu et al. (2023); Panaganti & Kalathil (2022); Yang & Wang (2019). *In fact, finding the sample complexity guarantee without the simulator is still an open question even for the unconstrained robust unconstrained MDP.*

Learning Goal: Since we do not know the uncertainty set, we cannot obtain an optimal policy from the beginning. Rather, the goal is to obtain a policy $\hat{\pi}$ such that for a given $\epsilon > 0$, using N_{tot} samples or queries from the generative model such that $\text{Sub} - \text{Opt}(\hat{\pi}) \leq \epsilon$, and $\text{Violation}(\hat{\pi}) \leq \epsilon$ (see

Definition 1). Unlike the unconstrained robust MDP, one needs to ensure that both the violation and the sub-optimality gap are small.

Robust Bellman Consistency equation: Directly applying the result from Iyengar (2005), we have for any π, for $j = r, g$, and for all h, and s,

$$V^{\pi}_{j,h}(s) = \sum_{a} \pi_h(a|s)[j_h(s,a) + L_{\mathcal{P}_{h,s,a}} V^{\pi}_{j,h+1}] \tag{6}$$

where $L_{\mathcal{P}_{h,s,a}} V = \inf\{PV : P \in \mathcal{P}_{h,s,a}\}$.

3 Solution Methodology

In this section, we first discuss why one cannot directly apply the primal-dual methods popular for solving the non-robust CMDP to the robust CMDP. Subsequently, we describe our novel approach.

Why existing primal-dual methods for non robust CMDP do not work?:

In the non-robust CMDP, one solve for the Lagrangian

$$\min_{\lambda \geq 0} \max_{\pi} V^{\pi,P}_{r}(x) + \lambda(V^{\pi,P}_{g}(x) - b) \tag{7}$$

It has been shown that strong duality holds if a strict feasible policy (a.k.a. Slater's condition) Paternain et al. (2019) exists. Thus, one can solve in the dual-domain. For non-robust CMDP problems, Vaswani et al. (2022); Ghosh et al. (2022) demonstrate that primal-dual based approaches can in fact achieve a feasible and an ϵ-sub optimal policy using $\tilde{O}(1/\epsilon^2)$ interactions with the environment Ghosh et al. (2022) or generative model Vaswani et al. (2022) using the strong duality result. The key to obtain strong duality result in Paternain et al. (2019) is that the state-action occupancy measure $d^{\pi,P}_h(s,a)$ for a given transition probability measure P is convex. In particular, consider π_1, and π_2 with the corresponding state-action occupancy measures $d^{\pi,P}_{1,h}$ and $d^{\pi,P}_{2,h}$, then there exists π' such that $(1-\lambda)d^{\pi,P}_{1,h}(s,a) + \lambda d^{\pi,P}_{2,h}(s,a) = d^{\pi',P}_{h}(s,a)$ for a $\lambda \in (0,1)$. All the existing primal-dual methods Ghosh et al. (2022); Wei et al. (2022); Ding et al. (2020); Efroni et al. (2020) uses such strong duality result to obtain sample complexity bound. However, the robust state-action occupancy measure is not convex as shown in Lemma 1 in Wang et al. (2022). Intuitively, the worst-case transition model depends on the policy, hence, the convexity is not assured. Thus, the robust CMDP may not admit strong duality even if there exists a strictly feasible policy. Hence, the traditional primal-dual based algorithms will be unable to obtain the sample complexity guarantees.

Second, in (7), for a given λ, one can solve for the optimal using the standard Dynamic programming approach as it would be a simply unconstrained problem with modified per-step reward $r_h + \lambda g_h$. Hence, the standard value-based approach can be applied for the composite value function. For unconstrained robust MDP, one can apply robust value iteration to solve for the policy Iyengar (2005). One might be curios why not apply the robust value iteration to solve for π for the Lagrangian $V^{\pi}_{r}(x) + \lambda(V^{\pi}_{g}(x) - b)$ for a given π similar to the non-robust CMDP. *However, this won't work as the worst case model for the reward value function and the worst-case model for the utility value function can be different.* In particular, $L_{\mathcal{P}_{h,s,a}} V^{\pi}_{r,h} + L_{\mathcal{P}_{h,s,a}} V^{\pi}_{g,h} \neq L_{\mathcal{P}_{h,s,a}}[V^{\pi}_{r,h} + V^{\pi}_{g,h}]$ because of the non-linearity unlike the non-robust CMDP. Thus, one cannot apply the robust value iteration to the combined value function to solve for the lagrangian $V^{\pi}_{r}(x) + \lambda(V^{\pi}_{g}(x) - b)$ for the robust CMDP even for a given fixed λ.

3.1 Rationale behind Our proposed approach

Since the traditional primal-dual algorithm will not work, we transform the problem into the following *rectified* penalty form:

$$\max_{\pi} \min_{P \in \mathcal{P}} V^{\pi,P}_{r}(x) - \lambda(b - \min_{P \in \mathcal{P}} V^{\pi,P}_{g}(x))_{+} \tag{8}$$

where $(x)_+ = \max\{x, 0\}$. Note that only when $\min_P V_g^{\pi,P}(x) < b$ (i.e., the policy does not satisfy the constraint), then it adds a penalty. Note that solving (8) might be simpler compared to solving a constrained problem which is not convex. Such a rectified penalty is inspired from Guo et al. (2022; 2023). However, they did not consider MDP setups. The advantage of the above formulation is captured below:

Lemma 1. *If $\lambda = 2H/\epsilon$, the optimal solution $\hat{\pi}$ of (8) will have a violation of at most ϵ, i.e., $(b - \min_{P\in\mathcal{P}} V_g^{\hat{\pi},P}(x)) \leq \epsilon$.*

The above result shows that the violation is bounded by $O(\epsilon)$ by a proper choice of λ. The above lemma also implies the following:

Corollary 1. *Fix $\lambda = 2H/\epsilon$. If for any infeasible policy π, $(\min_{P\in\mathcal{P}} V_g^{\pi,P}(x) - b) < \epsilon$, then π^* (optimal solution of (1) is also the optimal solution of (8).*

Remark 1. *Corollary 1 shows that if one makes $\lambda \to \infty$, the optimal solution of (8) coincides with (1). The proofs of Lemma 1 and Corollary 1 are in Appendix A.*

Remark 2. *Corollary 1 also states that if there is a gap of ϵ in the value functions for the feasible and infeasible policies, i.e, $\min_P V_g^{\pi,P} < b - \epsilon$ for any infeasible π, then solving (8) is equivalent to solving (1) for a finite value, $\lambda = 2H/\epsilon$. Thus, if there is a gap between the value functions for the feasible policies, and infeasible policies then solving (8) is enough.*

Algorithm 1 Robust Safe Reinforcement Learning Algorithm

1: **Input:** Uncertainty radius ρ, Performance bound ϵ, $\lambda = 6H/\epsilon$.
2: **Initialization:** $\hat{V}_{r,H+1}(s) = 0$, $\hat{V}_{g,H+1}(s) = 0$, $\forall s$.
3: **for** steps $h = H, H-1, \ldots, 1$ **do**
4: Collect N (value is in Section 4) samples for each state-action pair $(s, a) \in S \times A$.
5: Compute the empirical uncertainty set for all $(s, a) \in S \times A$ according to (9).
6: $\hat{\pi} = \arg\max_\pi \hat{V}_r^\pi(x) - \lambda(b - \epsilon/3 - \hat{V}_g^\pi(x))_+$ (the computation of $\hat{V}_r^\pi(x)$, and $\hat{V}_g^\pi(x)$ for a given π is obtained according to (10) in backward induction manner).

3.2 Handling unknown uncertain transition probability kernel set

We now discuss how we address the other challenge where the agent does not know the nominal transition probability model P^0, and it only has access to a simulator where the next states are drawn from the nominal model P^0 for a given state-action pair. The agent needs to learn efficiently to optimize the sample complexity.

We consider that the agent is estimating the transition probability model, $\hat{P}^0$, of the nominal model. Formally, in Algorithm 1 we generate N samples from each state-action pair at step h. We denote $N_h(s, a, s')$ as the number of times the state-action pair (s, a) transitions to s', then $\hat{P}_h^0(s'|s, a) = N_h(s, a, s')/N$. We then build the empirical uncertainty set as

$$\hat{\mathcal{P}} = \bigotimes \hat{\mathcal{P}}_{h,s,a}, \quad \hat{\mathcal{P}}_{h,s,a} = \{P \in \Delta(S) | D(P, \hat{P}^0) \leq \rho\} \tag{9}$$

After collecting $N_{tot} = NH|S||A|$ samples, we have an estimate for the value function for estimated model $\hat{P}$ as $\hat{V}_{j,h}^{\pi,\hat{P}}(s)$, and the robust empirical value function $\hat{V}_{j,h}^\pi(x) = \min_{\hat{P}\in\hat{\mathcal{P}}} \hat{V}_{j,h}^{\pi,P}(s)$ for any π, s, and h from the robust Bellman consistent equation for $j = r, g$ which we obtain recursively as $\hat{V}_{j,h+1}(s) = 0$ for all s. In particular, similar to (6), we compute

$$\hat{V}_{j,h}^\pi(s) = \sum_a \pi(a|s)[j_h(s, a) + L_{\hat{P}_{h,s,a}} \hat{V}_{j,h+1}^\pi] \tag{10}$$

where $L_{\hat{P}_{h,s,a}} \hat{V} = \inf\{PV : P \in \hat{P}_{h,s,a}\}$. For different uncertainty set, one can do an exhaustive search for finite state-space to compute $L_{\hat{P}_{h,s,a}} \hat{V}$. However, we will discuss later how to efficiently compute $L_{\hat{P}_{h,s,a}} \hat{V}$ for different uncertainty sets.

Rectified penalty for estimated model: Note that π^* is feasible for the original problem (1). Since we do not know the nominal model rather we are learning, there will be an error in estimating $V_g^\pi(x)$ for a policy π. In particular, the estimated value function for the optimal policy $\hat{V}_g^{\pi^*}(x) < b$ even though $V_g^{\pi^*}(x) \geq b$ because of the error. If we do not account for that, while solving for (8), we will get a conservative policy and will obtain a poor sub-optimality gap. Rather, we need to relax the constraint a little to maintain the balance.

In other words, we have to bound the error between $V_g^{\pi^*}(x)$ and $\hat{V}_g^{\pi^*}(x)$. In the standard non-robust case one can use the standard concentration inequality of Hoeffding to bound the error between the empirical value and the actual value. In particular, for the non-robust case, we have

$$V_j^{\pi,P^0}(x) - \hat{V}_j^{\pi,\hat{P}^0}(x) \leq H \max_{h,s,a} |P^0\hat{V} - \hat{P}^0\hat{V}|.$$

One can bound the error in the above using standard concentration inequality since $\mathbb{E}[\hat{P}^0] = P^0$.

However, in the robust case, it is more challenging as we cannot apply the standard concentration bound. Rather, we obtain in Appendix B–

Lemma 2. *For a given π,*

$$V_j^\pi(x) - \hat{V}_j^\pi(x) \leq H \max_{h,s,a} |L_{\mathcal{P}_{h,s,a}}\hat{V}_j^\pi - L_{\hat{\mathcal{P}}_{h,s,a}}\hat{V}_j^\pi| \tag{11}$$

Unlike the non-robust case, because of the non-linearity, $\mathbb{E}[L_{\hat{\mathcal{P}}}\hat{V}] \neq L_{\mathcal{P}}\hat{V}$. Hence, the standard concentration inequalities won't work.

In order to (11) we use the same technique as proposed in Xu et al. (2023). In particular, we obtain a dual representation for $L_{\mathcal{P}}V$ to obtain a tighter upper bound for (11). In the following, we describe the procedure for the *total variational distance metric*. One can easily extend it to other metrics from the results in Xu et al. (2023) and in Appendix F.1.

Total Variational Distance: For total variational distance metric, from Proposition 2 we have

Proposition 1. *For any $(s,a,h) \in S \times A \times [H]$, and $j = r, g$, $\theta, \delta \in (0,1)$, and $\rho \in (0,1]$, we have with probability $1-\delta$,*

$$H \max_{h,s,a} |L_{P^{TV}_{h,s,a}}\hat{V}_{j,h} - L_{\hat{P}^{TV}_{h,s,a}}\hat{V}_{j,h}| \leq 2\theta H + H\sqrt{H^2 \log(4H^2|S||A|/\delta\theta)/(2N)} \tag{12}$$

where we use the θ-covering number for the value function.

Using $\theta = \epsilon/(12H)$, and

$$N = \frac{18H^4 \log(48H^3|S||A|/\epsilon\delta)}{\epsilon^2} \tag{13}$$

in Lemma 2, we obtain (see Appendix B)

Lemma 3. *For a given policy π, for $j = r, g$ with probability $1-2\delta$ when N is set according to (13),*

$$|V_j^\pi(x) - \hat{V}_j^\pi(x)| \leq \epsilon/3$$

Relaxed Problem: Hence, in order to ensure that the optimal policy π^* of the original robust CMDP (cf.(1)) problem, we consider the following in terms of the estimated value functions.

$$\max_\pi \hat{V}_r^\pi(x) \quad \text{s.t } \hat{V}_g^\pi(x) \geq b - \epsilon/3. \tag{14}$$

when N is set according to (13) in Algorithm 1 for total variation uncertainty set. By Lemma 3, it ensures that the optimal policy of the original problem (1) is feasible to the (14) with high probability accounting for the estimated value function. Note that even though we consider the relaxed problem, we still obtain the violation bound of ϵ by proper choice of N. *In particular, we select N such that the constraint becomes $\hat{V}_g^\pi \geq b - \epsilon/3$ for a given $\epsilon > 0$.* For different uncertainty sets, the values of N to bound by $\epsilon/3$ are characterized in Appendix D.

3.3 Optimization Oracle

As we described, standard primal-dual approach won't work. Based on the insights obtained in Lemma 1 for the rectified penalty objective (cf.(8)), our algorithm 1 *assumes* the existence of the optimization oracle (see line 6) that finds the following policy

$$\max_{\pi} \min_{\hat{P}\in\hat{\mathcal{P}}} \hat{V}_r^{\pi,P}(x) - \lambda \left(b - \epsilon/3 - \min_{\hat{P}\in\hat{\mathcal{P}}} \hat{V}_g^{\pi,P}(x) \right)_+ \tag{15}$$

As we described in Lemma 1, for $\lambda = 6H/\epsilon$, one can guarantee that it achieves a policy with violation of at most $\epsilon/3$ for (14). We use the above bound for our analysis.

One can use policy gradient or policy optimization approach to find such a policy (e.g., using neural network). We also discuss on how to develop such a policy-optimization oracle in Section 5. However, characterization of the sample complexity analysis without such an oracle has been left for the future.

4 Main Results

We now state the main result of our paper. We prove it in Appendix C.

Theorem 1. *For total variation distance uncertainty set, after $N_{tot} \geq N_{TV}$ samples, where*

$$N_{TV} = \frac{18H^5|S||A|}{\epsilon^2} \log\left(\frac{48H^3|S||A|}{\epsilon\delta}\right) \tag{16}$$

Algorithm 1 returns the policy $\hat{\pi}$ such that with probability $1-3\delta$, $\text{Sub}-\text{Opt}(\hat{\pi}) \leq \epsilon$, and $\text{Violation}(\hat{\pi}) \leq \epsilon$.

The result indicates that one needs $\tilde{\mathcal{O}}(H^5|S||A|/\epsilon^2)$ samples to bound both the sub − opt and Violation by ϵ. This is the *first* such sample complexity result result for the robust CMDP. Note that the bound matches that of known bounds for the robust unconstrained MDP Xu et al. (2023). The lower bound $\Omega(H^4|S||A|/\epsilon^2)$ for the unconstrained non-robust MDP, hence, we are off by a factor of H, however, it is optimal (nearly) in terms of ϵ. Note that such a gap also exists for the unconstrained case as well.

Remark 3. ***Exact Feasibility****: Note that we can obtain exact feasibility for the scenario where $V_g^{\pi}(x) < b - \epsilon$ for any infeasible π. Also, note that we can achieve exact feasibility if $V_g^{\pi^*}(x) \geq b + \epsilon$ (i.e., the optimal policy is not in the boundary), by simply replacing b with $b + \epsilon$ in the algorithm.*

5 Discussions, Limitations, and Future Works

We provide the sample complexity of $\tilde{\mathcal{O}}(H^5|S||A|/\epsilon^2)$ guarantee using a generative model and an optimization oracle. We characterize the number of samples required to bound the sub-optimality and violation gaps by ϵ amount for various popular distance metrics.

Dependency on the oracle: We can develop a practical version of our approach (while compromising on the sample complexity) based on the double loop approach proposed by Wang et al. (2023) for unconstrained robust MDP–in the inner loop, the algorithm seeks to obtain the worst-case models for the reward value function and the utility value function for a given policy. In the outer loop, the policy parameter is updated using the policy gradient approach. Note that taking the gradient with respect to π in (8) is easy. The characterization of the sample complexity result for robust CMDP without the policy optimization oracle constitutes an important future research work.

We consider a tabular case. The characterization of the sample complexity guarantee for large state-space (possibly infinite) also constitutes an interesting future research direction. Reducing the dependence on H also remains an open question. Finally, our algorithm is model-based, characterizing the sample complexity guarantee using model-free approach is an important future research direction.

References

Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC press, 1999.

Archana Bura, Aria HasanzadeZonuzy, Dileep Kalathil, Srinivas Shakkottai, and Jean-Francois Chamberland. Dope: Doubly optimistic and pessimistic exploration for safe reinforcement learning. *Advances in Neural Information Processing Systems*, 35:1047–1059, 2022.

Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo R Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. In *NeurIPS*, 2020.

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 3304–3312. PMLR, 2021.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Alejandro Ribeiro. Last-iterate convergent policy gradient primal-dual methods for constrained mdps. *arXiv preprint arXiv:2306.11700*, 2023.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

Yingjie Fei, Zhuoran Yang, Yudong Chen, Zhaoran Wang, and Qiaomin Xie. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. *Advances in Neural Information Processing Systems*, 33:22384–22395, 2020.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91: 325–349, 2013.

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Provably efficient model-free constrained rl with linear function approximation. *Advances in Neural Information Processing Systems*, 35:13303–13315, 2022.

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Towards achieving sub-linear regret and hard constraint violation in model-free rl. In *International Conference on Artificial Intelligence and Statistics*, pp. 1054–1062. PMLR, 2024.

Hengquan Guo, Xin Liu, Honghao Wei, and Lei Ying. Online convex optimization with hard constraints: Towards the best of two worlds and beyond. *Advances in Neural Information Processing Systems*, 35:36426–36439, 2022.

Hengquan Guo, Zhu Qi, and Xin Liu. Rectified pessimistic-optimistic learning for stochastic continuum-armed bandit with constraints. In *Learning for Dynamics and Control Conference*, pp. 1333–1344. PMLR, 2023.

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

Dileep Kalathil, Vivek S Borkar, and Rahul Jain. Empirical q-value iteration. *Stochastic Systems*, 11 (1):1–18, 2021.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Xin Liu, Bin Li, Pengyi Shi, and Lei Ying. An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. *Advances in Neural Information Processing Systems*, 34, 2021.

Xiaoteng Ma, Zhipeng Liang, Jose Blanchet, Mingwen Liu, Li Xia, Jiheng Zhang, Qianchuan Zhao, and Zhengyuan Zhou. Distributionally robust offline reinforcement learning with linear function approximation. *arXiv preprint arXiv:2209.06620*, 2022.

Daniel J Mankowitz, Dan A Calian, Rae Jeong, Cosmin Paduraru, Nicolas Heess, Sumanth Dathathri, Martin Riedmiller, and Timothy Mann. Robust constrained reinforcement learning for continuous control with model misspecification. *arXiv preprint arXiv:2010.10644*, 2020.

Arnab Nilim and Laurent Ghaoui. Robustness in markov decision problems with uncertain transition matrices. *Advances in neural information processing systems*, 16, 2003.

Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pp. 9582–9602. PMLR, 2022.

Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. *Advances in neural information processing systems*, 35:32211–32224, 2022.

Santiago Paternain, Luiz Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. *Advances in Neural Information Processing Systems*, 32, 2019.

Reazul Hasan Russel, Mouhacine Benosman, and Jeroen Van Baar. Robust constrained-mdps: Soft-constrained robust policy optimization under model uncertainty. *arXiv preprint arXiv:2010.04870*, 2020.

Zhongchang Sun, Sihong He, Fei Miao, and Shaofeng Zou. Constrained reinforcement learning under model mismatch. *arXiv preprint arXiv:2405.01327*, 2024.

Niko Sünderhauf, Oliver Brock, Walter Scheirer, Raia Hadsell, Dieter Fox, Jürgen Leitner, Ben Upcroft, Pieter Abbeel, Wolfram Burgard, Michael Milford, et al. The limits and potentials of deep learning for robotics. *The International journal of robotics research*, 37(4-5):405–420, 2018.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017.

Sharan Vaswani, Lin F Yang, and Csaba Szepesvári. Near-optimal sample complexity bounds for constrained mdps. *arXiv preprint arXiv:2206.06270*, 2022.

Qiuhao Wang, Chin Pang Ho, and Marek Petrik. Policy gradient in robust mdps with global convergence guarantee. In *International Conference on Machine Learning*, pp. 35763–35797. PMLR, 2023.

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

Yue Wang and Shaofeng Zou. Policy gradient method for robust reinforcement learning. In *International conference on machine learning*, pp. 23484–23526. PMLR, 2022.

Yue Wang, Fei Miao, and Shaofeng Zou. Robust constrained reinforcement learning. *arXiv preprint arXiv:2209.06866*, 2022.

Honghao Wei, Xin Liu, and Lei Ying. Triple-q: A model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3274–3307. PMLR, 2022.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Tengyu Xu, Yingbin Liang, and Guanghui Lan. Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, pp. 11480–11491. PMLR, 2021.

Zaiyan Xu, Kishan Panaganti, and Dileep Kalathil. Improved sample complexity bounds for distributionally robust reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 9728–9754. PMLR, 2023.

Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pp. 6995–7004. PMLR, 2019.

Zhengqing Zhou, Zhengyuan Zhou, Qinxun Bai, Linhai Qiu, Jose Blanchet, and Peter Glynn. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3331–3339. PMLR, 2021.

Organization of Appendix: We prove Lemma 1 and Corollary 1 in Appendix A. We prove Lemmas 2 and 3 in Appendix B. Subsequently, We prove Theorem 1 in Appendix C. We state the other uncertainty sets and state the complexity results in Appendix D. We prove Theorems 3, 2, and 4 in Appendix E. Finally, we state some important results in Appendix F which we use for proving results.

A Proof of Lemma 1 and Corollary 1

A.1 Proof of Lemma 1

Proof. We prove this by contradiction. Suppose that the statement is not true.

Consider π^* the optimal solution of the original problem (1). Since this is feasible, then $(b-V_g^{\pi^*}(x))_+ = 0$, then

$$V_r^{\pi^*}(x) - \lambda(b - V_g^{\pi^*})_+ \geq -H \tag{17}$$

as $|r_h(x,a)| \leq 1$. For any π (including the optimal policy $\hat{\pi}$) of (8) ,we have

$$V_r^{\pi} - \lambda(b - V_g^{\pi})_+ \leq H - \lambda(b - V_g^{\pi})_+ \tag{18}$$

For the optimal policy $\hat{\pi}$, we have $(b - \min_{P\in\mathcal{P}} V_g^{\hat{\pi},P}(x)) > \epsilon$ (by contradiction). Hence, by (18), we have

$$V_r^{\hat{\pi}} - \lambda(b - V_g^{\hat{\pi}})_+ < H - (2H/\epsilon)\epsilon = -H$$

However, it contradicts (17) as π^* can achieve a better value for the objective in (8). □

A.2 Proof of Corollary 1

Proof. By Lemma 1 the optimal policy in (8) must have a constraint violation of at most ϵ. Now, since any infeasible policy now must satisfy the condition $(b - \min_{P\in\mathcal{P}} V_g^{\pi,P}(x)) < \epsilon$, it means that only the feasible policies of the original robust CMDP (1) can be optimal solution of (8). Hence,

$$\min_{P\in\mathcal{P}} V_r^{\hat{\pi},P} - \lambda(b - \min_{P\in\mathcal{P}} V_g^{\hat{\pi},P}(x))_+ \leq \min_{P\in\mathcal{P}} V_r^{\pi^*,P} - \lambda(b - \min_{P\in\mathcal{P}} V_g^{\pi^*,P}(x))_+$$

where we have used the fact that $(b - \min_{P\in\mathcal{P}} V_g^{\pi^*,P}(x))_+ = 0$, and $(b - \min_{P\in\mathcal{P}} V_g^{\hat{\pi},P}(x))_+ = 0$. Thus, the optimal solution of (1) is actually the optimal in (8). □

B Proof of Lemmas 2 and 3

B.1 Proof of Lemma 2

Proof. From (6) and (10) we have $\forall s$ and $j = r, g$,

$$\begin{aligned}\hat{V}^{\pi}_{j,h}(s) - V^{\pi}_{j,h}(s) &= \sum_a \pi(a|s)[j_h(s,a) + L_{\hat{\mathcal{P}}_h(s,a)}\hat{V}^{\pi}_{j,h+1}] - \sum_a \pi(a|s)[j_h(s,a) + L_{\mathcal{P}_h(s,a)}V^{\pi}_{j,h+1}] \\ &= \sum_a \pi(a|s)[L_{\hat{\mathcal{P}}_h(s,a)}\hat{V}^{\pi}_{j,h+1} - L_{\mathcal{P}_h(s,a)}V^{\pi}_{h+1}] \\ &= \sum_a \pi(a|s)[L_{\hat{\mathcal{P}}_h(s,a)}\hat{V}^{\pi}_{j,h+1} - L_{\mathcal{P}_h(s,a)}\hat{V}^{\pi}_{j,h+1} + L_{\mathcal{P}_h(s,a)}\hat{V}^{\pi}_{j,h+1} - L_{\mathcal{P}_h(s,a)}V^{\pi}_{h+1}] \\ &\leq \max_a |L_{\hat{\mathcal{P}}_h(s,a)}\hat{V}^{\pi}_{j,h+1} - L_{\mathcal{P}_h(s,a)}\hat{V}^{\pi}_{j,h+1}| + \max_a |L_{\mathcal{P}_h(s,a)}\hat{V}^{\pi}_{j,h+1} - L_{\mathcal{P}_h(s,a)}V^{\pi}_{j,h+1}|\end{aligned}$$

where we use Holder's inequality and $||\pi||_1 = 1$. Since the above holds for every s, thus, for any s

$$\hat{V}^{\pi}_{j,h}(s) - V^{\pi}_{j,h}(s) \leq \max_a |L_{\hat{\mathcal{P}}_h(s,a)}\hat{V}^{\pi}_{j,h+1} - L_{\mathcal{P}_h(s,a)}\hat{V}^{\pi}_{j,h+1}| + \max_a |L_{\mathcal{P}_h(s,a)}\hat{V}^{\pi}_{j,h+1} - L_{\mathcal{P}_h(s,a)}V^{\pi}_{j,h+1}|$$

Expanding recursively, we obtain

$$|\hat{V}^{\pi}_{j,1}(s) - V^{\pi}_{j,1}(s)| \leq \max_s \max_a \max_h H|L_{\mathcal{P}_h(s,a)}\hat{V}^{\pi}_{j,h+1} - L_{\hat{\mathcal{P}}_h(s,a)}\hat{V}^{\pi}_{j,h+1}| \tag{19}$$

□

B.2 Proof of Lemma 3

Note from Lemma 2 we obtain

$$|\hat{V}^{\pi}_{j,1}(s) - V^{\pi}_{j,1}(s)| \leq \max_s \max_a \max_h H|L_{\mathcal{P}_h(s,a)}\hat{V}^{\pi}_{j,h+1} - L_{\hat{\mathcal{P}}_h(s,a)}\hat{V}^{\pi}_{j,h+1}|$$

Now, from lemma 4 and applying union bound over (h, s, a) we obtain with probability $1 - \delta$

$$|\hat{V}^{\pi}_{j,1}(s) - V^{\pi}_{j,1}(s)| \leq \sqrt{\frac{H^4 \log(4H^2|S||A|/(\theta\delta))}{2N}} + 2\theta H \tag{20}$$

Taking $\theta = \epsilon/12H$, and $N = \dfrac{18H^4 \log(16H^3|S||A|/(\epsilon\delta))}{\epsilon^2}$, we obtain

$$|\hat{V}^{\pi}_{j,1}(s) - V^{\pi}_{j,1}(s)| \leq \epsilon/3 \tag{21}$$

C Proof of Theorem 1

Bounding the Sub-optimality bound: First, we observe the following

$$V^{\pi^*}_r(x) - V^{\hat{\pi}}_r(x) = \underbrace{V^*_r(x) - \hat{V}^{\pi^*}_r}_{T_1} + \underbrace{\hat{V}^{\pi^*}_r - \hat{V}^{\hat{\pi}}_r(x)}_{T_2} + \underbrace{\hat{V}^{\hat{\pi}}_r(x) - V^{\hat{\pi}}_r(x)}_{T_3} \tag{22}$$

Since $N = 18H^4 \log(48H^3|S||A|/(\epsilon\delta))/\epsilon^2$, by Lemma 3, T_1, and T_3 both can be bounded by $\epsilon/3$ with probability $1 - 2\delta$. We now bound T_2 which we describe next.

Since $\hat{\pi}$ is optimal of (15), thus, replacing the value of N there, we obtain

$$\hat{V}^{\pi^*}_r(x) - \lambda(b - \epsilon/3 - \hat{V}^{\pi^*}_g(x))_+ \leq \hat{V}^{\hat{\pi}}_r(x) - \lambda(b - \epsilon/3 - \hat{V}^{\hat{\pi}}_g(x))_+ \tag{23}$$

By Lemma 3, we have $(b - \epsilon/3 - \hat{V}^{\pi^*}_g(x)) \leq 0$ with probability $1 - \delta$. Hence,

$$\hat{V}^{\pi^*}_r(x) - \hat{V}^{\hat{\pi}}_r(x) \leq -\lambda(b - \epsilon/3 - \hat{V}^{\hat{\pi}}_g(x))_+ \leq 0 \tag{24}$$

where the inequality follows from the fact that $(x)_+ \geq 0$.

Hence, $T_2 \leq 0$ with probability $1-\delta$. Thus, combining all, we have with probability $1-3\delta$,

$$V_r^*(x) - V_r^{\hat{\pi}}(x) \leq \epsilon/3 + \epsilon/3 = 2\epsilon/3 \leq \epsilon.$$

Violation Bound: Now, we show the violation bound. We decompose in the following way–

$$(b - V_g^{\hat{\pi}}(x)) = \underbrace{(b - \epsilon/3 - \hat{V}^{\hat{\pi}}(x)}_{T_4} + \epsilon/3 + \underbrace{\hat{V}^{\hat{\pi}}(x) - V_g^{\hat{\pi}}(x))}_{T_5} \tag{25}$$

We bound T_5 using Lemma 3 by $\epsilon/3$. Note that using the value of N, we obtain, $\sqrt{\frac{3H^4 \log(16H^3|S||A|)}{N\delta}} \leq \epsilon/3$. In order to bound T_4 we use Lemma 1 to obtain

$$(b - \epsilon/3 - \hat{V}_g^{\hat{\pi}}(x)) \leq \epsilon/3 \tag{26}$$

when $\lambda = 6H/\epsilon$.

Hence, combining all, we obtain

$$(b - V_g^{\hat{\pi}}(x)) \leq \epsilon/3 + \epsilon/3 + \epsilon/3 = \epsilon$$

D Results for other Uncertainty Sets

1. **Chi-squared uncertainty set**: Let $\mathcal{P}^{\chi} = \bigotimes_{(h,s,a)\in[H]\times S\times A} \mathcal{P}^{\chi}_{h,s,a}$ be the uncertainty set defined in (4) with chi-squared distance Xu et al. (2023)

$$D_{\chi}(P, P^0_{h,s,a}) = \sum_{s'} \frac{(P(s') - P^0_{h,s,a}(s'))^2}{P^0_{h,s,a}(s')} \tag{27}$$

2. **KL-uncertainty set**: Let $\mathcal{P}^{KL} = \bigotimes_{(h,s,a)\in[H]\times S\times A} \mathcal{P}^{KL}_{h,s,a}$ be the uncertainty set defined in (4) with KL-divergence metric Kullback & Leibler (1951)

$$D_{KL}(P, P^0_{h,s,a}) = \sum_{s'} P(s') \log\left(\frac{P(s')}{P^0_{h,s,a}(s')}\right) \tag{28}$$

3. **Wasserstein uncertainty set**: Let $\mathcal{P}^{W} = \bigotimes_{(h,s,a)\in[H]\times S\times A} \mathcal{P}^{W}_{h,s,a}$ be the uncertainty set defined in (4) with the Wasserstein distance metric Xu et al. (2023)

$$D_{W}(P, P^0_{h,s,a}) = \inf_{\nu \in m(P, P^0_{h,s,a})} \int d^p(x,y) d\nu(x,y) \tag{29}$$

where the integration is over all $(x,y) \in S \times S, p \in [1,\infty)$, and $m(P, P^0_{h,s,a})$ denote all probability measures on $S \times S$ with marginals P and $P^0_{h,s,a}$. In addition, we set $B_p = \max_{s,s'} d^p(s,s')$.

We now state the results for the other distance metrics

Theorem 2. *For Chi-squared uncertainty set, if the total number of samples $N_{tot} \geq N_{\chi}$ where*

$$N_{\chi} = \frac{288 C_\rho^4 H^5 |S||A|}{(C_\rho - 1)^2 \epsilon^2} \log\left(\frac{2H|S||A|\left(1 + \frac{12 C_\rho H^2}{\epsilon(C_\rho - 1)}\right)}{\delta}\right) \tag{30}$$

and where $C_\rho = \sqrt{1+\rho}$, then, the policy $\hat{\pi}$ returned by Algorithm 1 satisfies with probability $1-3\delta$, $\text{Sub} - \text{Opt}(\hat{\pi}) \leq \epsilon$, *and* $\text{Violation}(\hat{\pi}) \leq \epsilon$.

Theorem 2 shows that the sample complexity result is $\tilde{\mathcal{O}}(H^5|S||A|/\epsilon^2)$ which is the same as the unconstrained case Xu et al. (2023). Remark 3 is also applicable here.

Theorem 3. *For the KL uncertainty set, if the total number of samples $N_{tot} \geq N_{KL}$ where*

$$N_{KL} = \frac{4.5\exp(3H/\zeta)H^5|S||A|}{\rho^2\epsilon^2}\log\left(\frac{8H|S||A|}{\zeta\delta}\right) \tag{31}$$

where ζ is the problem-dependent parameter, and independent N_{KL}, then the policy $\hat{\pi}$ returned by Algorithm 1 satisfies with probability $1-3\delta$ $\text{Sub}-\text{opt}(\hat{\pi}) \leq \epsilon$, and $\text{Violation}(\hat{\pi}) \leq \epsilon$.

Note that here the sample complexity bound is $\tilde{\mathcal{O}}(\exp(H)H^5|S||A|/\epsilon^2)$. The exponential dependence is also observed in the unconstrained non-robust case as well Xu et al. (2023). Recently, Fei et al. (2020) shows that exponential dependency on H is unavoidable for the risk-sensitive MDP (similar to the MDP with KL uncertainty set). Remark 3 is also applicable here.

Theorem 4. *For the Wasserstein uncertainty set, if the total number of samples $N_{tot} \geq N_W$ where*

$$N_W = \frac{18H^5|S||A|(B_p+\rho^p)^2}{\rho^{2p}\epsilon^2}\log\left(\frac{24H^2|S||A|(HB_p+\max(H,\rho^p))}{\rho^p\delta\epsilon}\right) \tag{32}$$

then, with probability $1-3\delta$, the policy $\hat{\pi}$ returned by Algorithm 1 satisfies $\text{Sub}-\text{Opt}(\hat{\pi}) \leq \epsilon$, and $\text{Violation}(\hat{\pi}) \leq \epsilon$.

The sample complexity bound here is $\tilde{\mathcal{O}}((B_p+\rho^p)^2H^5|S||A|/\rho^{2p}\epsilon^2)$ which is the same as in the unconstrained case Xu et al. (2023). Remark 3 is also applicable here.

E Proofs of Theorems 2, 3, and 4

E.1 Proof of Theorem 2

By applying Lemma 2, 5, and union bound we obtain with probability $1-\delta$

$$|\hat{V}^{\pi}_{j,1}(s) - V^{\pi}_{j,1}(s)| \leq H\frac{\sqrt{2}C_\rho^2 H}{(C_\rho-1)^2\sqrt{N}}\left(\sqrt{\log\left(\frac{2H|S||A|(1+C_\rho H/(\theta(C_\rho-1)))}{\delta}\right)}+1\right)+2\theta H \tag{33}$$

Now, using $\theta = \epsilon/(12H)$, and

$$N = \frac{72C_\rho^4H^4}{(C_\rho-1)^2\epsilon^2}\left(\sqrt{\log\left(\frac{2H|S||A|(1+C_\rho 12H/(\epsilon(C_\rho-1)))}{\delta}\right)}+1\right)^2 \tag{34}$$

we have with probability $1-\delta$,

$$|\hat{V}^{\pi}_{j,1}(s) - V^{\pi}_{j,1}(s)| \leq \epsilon/3 \tag{35}$$

Using the identity $(a+b)^2 \leq 2a^2+2b^2$, we get the sample complexity result that

$$|\hat{V}^{\pi}_{j,1}(s) - V^{\pi}_{j,1}(s)| \leq \epsilon/3 \tag{36}$$

when

$$N = \frac{288C_\rho^4H^4}{(C_\rho-1)^2\epsilon^2}\log\left(\frac{2H|S||A|(1+C_\rho 12H/(\epsilon(C_\rho-1)))}{\delta}\right) \tag{37}$$

The rest of the proof is similar to the proof of Theorem 1. Hence, we omit it here.

E.2 Proof of Theorem 3

By applying Lemma 2, 6, and union bound we obtain with probability $1-\delta$

$$|\hat{V}^{\pi}_{j,1}(s) - V^{\pi}_{j,1}(s)| \leq \frac{H^2}{\rho}\exp(\theta H)\exp(H/\zeta)\sqrt{\frac{\log(4H|S||A|/\theta\zeta\delta)}{2N}} \tag{38}$$

Now, taking $\theta = 1/2$, and putting

$$N = \frac{4.5H^4\exp(3H/\zeta)\log(8H|S||A|/\zeta\delta)}{\rho^2\epsilon^2} \tag{39}$$

then

$$|\hat{V}^{\pi}_{j,1}(s) - V^{\pi}_{j,1}(s)| \leq \epsilon/3 \tag{40}$$

with probability $1-\delta$. The rest of the proof follows the same as in Theorem 1.

E.3 Proof of Theorem 4

By applying Lemma 2, 7, and union bound we obtain with probability $1-\delta$

$$|\hat{V}^{\pi}_{j,1}(s) - V^{\pi}_{j,1}(s)| \leq \frac{H^2(B_p+\rho^p)}{\rho^p}\sqrt{\frac{\log(H|S||A|\frac{HB_p+\max(H,\rho^p)}{\rho^p\theta\delta})}{2N}} + 2\theta H \tag{41}$$

Taking $\theta = \epsilon/(12H)$, and plugging the following value of N,

$$N = \frac{18H^4(B_p+\rho^p)^2}{\rho^{2p}\epsilon^2}\log(24H^2|S||A|\frac{2HB_p+2\max(H,\rho^p)}{\rho^p\epsilon\delta}) \tag{42}$$

then, we have with probability $1-\delta$,

$$|\hat{V}^{\pi}_{j,1}(s) - V^{\pi}_{j,1}(s)| \leq \epsilon/3 \tag{43}$$

The rest of the proof is the same as in the proof of Theorem 1. Hence, we omit it here.

F Technical Lemmas

Lemma 4. *Xu et al. (2023) For the total-variation uncertainty set, for any $V \in \mathcal{V}$, and fix any $(h,s,a) \in [H]\times S\times A$. For any $\theta,\delta \in (0,1)$, and $\rho > 0$, we have with probability $1-\delta$,*

$$|L_{\mathcal{P}_{h,s,a}}V - L_{\hat{\mathcal{P}}_{h,s,a}}V| \leq \sqrt{\frac{H^2\log(4H/(\theta\delta))}{2N}} + 2\theta \tag{44}$$

where N is the number of samples used to approximate $P^0_{h,s,a}$.

Note that θ is the covering number for the value function.

Lemma 5. *Xu et al. (2023)For Chi-squared uncertainty set, fix any $V \in \mathcal{V}$, and fix any $(h,s,a) \in [H]\times S\times A$. For any $\theta,\delta \in (0,1)$, and $\rho > 0$, we have with probability $1-\delta$,*

$$|L_{\mathcal{P}_{h,s,a}}V - L_{\hat{\mathcal{P}}_{h,s,a}}V| \leq \frac{\sqrt{2}C_\rho^2 H}{(C_\rho-1)^2\sqrt{N}}\sqrt{\log(\frac{2(1+C_\rho H/(\theta(C_\rho-1))}{\delta}) + 1)} \tag{45}$$

where N is the number of samples used to approximate $P^0_{h,s,a}$.

Lemma 6. *Xu et al. (2023) For KL-divergence uncertainty set, fix any $V \in \mathcal{V}$, and fix any $(h,s,a) \in [H] \times S \times A$. For any $\theta, \delta \in (0,1)$, and $\rho > 0$, we have with probability $1-\delta$,*

$$|L_{\mathcal{P}_{h,s,a}} V - L_{\hat{\mathcal{P}}_{h,s,a}} V| \leq \frac{H}{\rho} \exp(\theta H) \exp(H/\zeta) \sqrt{\frac{\log(4/\theta\zeta\delta)}{2N}} \tag{46}$$

where N is the number of samples used to approximate $P^0_{h,s,a}$. ζ problem dependent parameter and independent of N.

Lemma 7. *Xu et al. (2023) For Wasserstein distance uncertainty set, for any $V \in \mathcal{V}$, and fix any $(h,s,a) \in [H] \times S \times A$. For any $\theta, \delta \in (0,1)$, and $\rho > 0$, we have with probability $1-\delta$,*

$$|L_{\mathcal{P}_{h,s,a}} V - L_{\hat{\mathcal{P}}_{h,s,a}} V| \leq \frac{H(B_p + \rho^p)}{\rho^p} \sqrt{\frac{\log(\frac{2HB_p + 2\max(H, \rho^p)}{\rho^p \theta \delta})}{2N}} + 2\theta \tag{47}$$

where N is the number of samples used to approximate $P^0_{h,s,a}$.

F.1 Efficient computation of robust value function for different uncertainty metrics

In this section, we state the results from Xu et al. (2023) and Yang & Wang (2019) for equivalent dual representation for worst-case value function evaluation for different uncertainty sets.

Proposition 2. *For total variation distance metric, and any value function V_j, $j = r, g$*

$$L_{\mathcal{P}^{TV}_{h,s,a}} V_j(s) = - \inf_{\eta \in [0, 2H/\rho]} \mathbb{E}_{s' \sim P^0_h(\cdot|s,a)}[(\eta - V_j(s'))_+] + \rho(\eta - \inf_{s'} V_j(s'))_+ - \eta \tag{48}$$

Proposition 3. *For χ-squared uncertainty set, for any value function $V \in \mathcal{V}$, we have*

$$L_{\mathcal{P}^{\chi}_{h,s,a}} V = - \inf_{\eta \in \mathbb{R}} \{\sqrt{\rho + 1} \sqrt{\mathbb{E}_{s' \sim P^0_h(\cdot|s,a)} (V(s') - \eta)^2}\} - \eta \tag{49}$$

The expression within the infimum is convex in η, and thus can be efficiently solved.

Proposition 4. *For KL-divergence uncertainty set, and for any value function $V \in \mathcal{V}$, we have*

$$L_{\mathcal{P}^{KL}_{h,s,a}} V = - \inf_{\eta \in [0, H/\rho]} \left\{ \eta\rho + \log\left(\mathbb{E}_{s' \sim P^0_h(\cdot|s,a)} [\exp(-V(s')/\eta)] \right) \right\} \tag{50}$$

Again the expression within the infimum is convex, and thus, can be efficiently solved.

Proposition 5. *For Wasserstein uncertainty set, and for any $V \in \mathcal{V}$, we have*

$$L_{\mathcal{P}^{W}_{h,s,a}} V = - \inf_{\eta \in [0, H/\rho^p]} \left(\eta\rho^p - \mathbb{E}_{s' \sim P^0_{h,s,a}} [\inf_{s''} \{V(s'') + \eta d^p(s'', s')\}] \right) \tag{51}$$

Again it is a convex optimization problem.