The Annals of Statistics
2022, Vol. 50, No. 6, 3223–3248
https://doi.org/10.1214/22-AOS2225

TOWARD THEORETICAL UNDERSTANDINGS OF ROBUST MARKOV DECISION PROCESSES: SAMPLE COMPLEXITY AND ASYMPTOTICS

BY WENHAO YANG[1,a], LIANGYU ZHANG[1,b] AND ZHIHUA ZHANG[2,c]

[1]*Academy for Advanced Interdisciplinary Studies, Peking University,* [a]*yangwenhaosms@pku.edu.cn,* [b]*zhangliangyu@pku.edu.cn*

[2]*School of Mathematical Sciences, Peking University,* [c]*zhzhang@math.pku.edu.cn*

In this paper, we study the nonasymptotic and asymptotic performances of the optimal robust policy and value function of robust Markov Decision Processes (MDPs), where the optimal robust policy and value function are estimated from a generative model. While prior work focusing on nonasymptotic performances of robust MDPs is restricted in the setting of the KL uncertainty set and (s, a)-rectangular assumption, we improve their results and also consider other uncertainty sets, including the L_1 and χ^2 balls. Our results show that when we assume (s, a)-rectangular on uncertainty sets, the sample complexity is about $\widetilde{\mathcal{O}}(\frac{|\mathcal{S}|^2|\mathcal{A}|}{\varepsilon^2\rho^2(1-\gamma)^4})$. In addition, we extend our results from the (s, a)-rectangular assumption to the s-rectangular assumption. In this scenario, the sample complexity varies with the choice of uncertainty sets and is generally larger than the case under the (s, a)-rectangular assumption. Moreover, we also show that the optimal robust value function is asymptotically normal with a typical rate $\sqrt{n}$ under the (s, a) and s-rectangular assumptions from both theoretical and empirical perspectives.

1. Introduction. Reinforcement Learning (RL) is a machine learning paradigm that addresses sequential decision-making problems in an unknown environment. Unlike the supervised learning scenario in which a labeled training dataset is provided, in RL the agent collects information by interacting with the environment through a course of actions. In addition to its success in empirical performance [27, 51, 52, 67], several works [13, 34, 35] provide insightful and solid theoretical understandings of RL. RL is typically formulated as the Markov Decision Processes (MDPs) problem [61]. The difficulty of solving an MDP is primarily attributable to the inexact knowledge of the reward R and transition probability P. To address the challenge, an alternative approach resorts to offline methods, where the agent only has access to a given explorable dataset generated by a strategy. Many practical deep RL algorithms employ the offline method and achieve state-of-the-art success empirically [21, 46, 52]. However, it often takes incredibly large datasets to make modern RL algorithms work. The matter of large sample size greatly hinders the application of RL in areas like policy-making, finance, and healthcare, where it is extremely expensive or even impossible to acquire such a large amount of data. Recently, there are many works focusing on sample efficiency of offline RL from a theoretical perspective. Some prior works have provided solid results on model-free offline methods [2, 8, 14] while others have considered model-based approaches [66, 78, 82, 83]. Through these theoretical efforts, sample-efficiently learning a near-optimal policy can be guaranteed, that is, the sample complexity is polynomial in parameters of the underlying MDPs.

In reality, sometimes the environment used to generate the offline dataset may be different from the real-world MDPs, resulting in suboptimal performance of the policy obtained by

Received November 2021; revised July 2022.
MSC2020 subject classifications. Primary 62C05, 62F12; secondary 68Q32.
Key words and phrases. Model-based reinforcement learning, robust MDPs, distributional robustness, f-divergence set.

RL algorithms. A well-known example is *the sim-to-real gap* [57, 84], which suggests that an RL-based robot controller trained in a simulated environment may perform poorly in the real world. A similar phenomenon also occurs in application scenarios such as healthcare and finance problems. For example, we may seek a dynamic treatment regime that would be deployed in hospital A using RL algorithms. However, the only available dataset is collected in hospital B. Naively performing RL algorithms with the given dataset and deploying the resulting regime in hospital A may cause bad outcomes. In addition, Mannor et al. [50] also showed that the value function might be sensitive to estimation errors of reward and transition probability, which means a small perturbation of reward and transition probability could incur a significant change in the value function. Then, robust MDPs [32, 55] have been proposed to handle these issues, where the transition probability is allowed to take values in an uncertainty set (or ambiguity set). In this way, the solution of robust MDPs is less sensitive to model estimation errors with a properly chosen uncertainty set $\widehat{\mathcal{P}}$.

In order to solve the robust MDP problem efficiently, one commonly makes the assumption that the uncertainty set $\widehat{\mathcal{P}}$ is either (s,a)-rectangular or s-rectangular [32, 55, 74], which stand for the transition probability P taking values independently for each state-action (s,a) pair or each state $s \in \mathcal{S}$, respectively. Compared with (s,a)-rectangular assumption, s-rectangular is a more general assumption to alleviate conservative policies and can provide stronger robustness guarantees [74]. Without these two assumptions, Wiesemann, Kuhn and Rustem [74] proved that solving robust MDPs could be NP-hard. However, under the (s,a)-rectangular or s-rectangular assumptions, the near-optimal robust policy and value function can be obtained efficiently. With these assumptions, Iyengar [32] and Nilim and El Ghaoui [55] proposed multiple choices of uncertainty sets under rectangular assumptions mentioned above, all of which are specific cases of f-divergence balls located around the estimated transition probability, including the L_1 distance, χ^2 and KL divergence balls. The most widely studied case is the so-called L_1 uncertainty set [3, 23, 31, 60] because it can be solved by the powerful linear programming methods.

In recent years, many works [25, 31, 47] have come up with efficient algorithms to solve robust MDPs, obtaining the optimal robust policy and value function. However, little theory has been developed on the statistical performances of the optimal robust policy and value function. Specifically, two core questions remain open: (a) How many samples are sufficient to guarantee the accuracy of the robust estimators? (b) Is it possible to make statistical inferences from the robust estimators? In this paper, we figure out both the finite-sample and asymptotic performances of the optimal robust policy and value function in different scenarios and answer these questions conclusively. Specifically, our nonasymptotic results in Sect. 3 show that sample-efficient reinforcement learning is possible in robust MDPs, which breaks the misconception that robust MDPs are exponentially hard in terms of effective horizon $(1-\gamma)^{-1}$ [85]. And our asymptotic results in Sect. 4 allow us to make statistical inferences from the robust estimators.

1.1. *Contributions*. Let $V_r^\pi(\mu)$ be the robust value function of policy π under uncertainty set $\mathcal{P}$ (unknown) and initial state distribution μ, and $\widehat{V}_r^\pi(\mu)$ be its empirical version under estimated uncertainty set $\widehat{\mathcal{P}}$. We denote by $\widehat{\pi}^* \in \operatorname{argmax}_\pi \widehat{V}_r^\pi(\mu)$ the optimal robust policy, and by $\widehat{V}_r^*(\mu) := \max_\pi \widehat{V}_r^\pi(\mu)$ the optimal robust value function. Rather than providing a new efficient algorithm to solve robust MDPs, we take efforts to study the statistical performances of optimal robust value function $\widehat{V}_r^*(\mu)$ and robust policy $\widehat{\pi}^*$ from both finite-sample and asymptotic perspectives. We mainly consider the frequently used data generating approach (i.e., generative models), from which we are able to estimate the transition probability $\widehat{P}$. Moreover, we consider three different uncertainty sets $\mathcal{P}$: L_1, χ^2, and KL balls under both (s,a) and s-rectangular assumptions, which are frequently applied in the field of

TABLE 1

The sample complexity of achieving ε deviation bound (1) in the generative model setting (Theorem 3.1). Here $|\mathcal{S}|$ and $|\mathcal{A}|$ are the sizes of the state space and action space, $\gamma \in (0,1)$ is a discount factor, ρ represents the size of uncertainty set in Examples 2.1 and 2.2, and $\underline{p} = \min_{P^*(s'|s,a)>0} P^*(s'|s,a)$

Uncertainty set	(s,a)-rectangular (Theorem 3.1)	s-rectangular (Theorem 3.2)
L_1	$\widetilde{\mathcal{O}}(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert(2+\rho)^2}{\varepsilon^2\rho^2(1-\gamma)^4})$	$\widetilde{\mathcal{O}}(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert^2(2+\rho)^2}{\varepsilon^2\rho^2(1-\gamma)^4})$
χ^2	$\widetilde{\mathcal{O}}(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert(1+\rho)^2}{\varepsilon^2(\sqrt{1+\rho}-1)^2(1-\gamma)^4})$	$\widetilde{\mathcal{O}}(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert^3(1+\rho)^2}{\varepsilon^2(\sqrt{1+\rho}-1)^2(1-\gamma)^4})$
KL	$\widetilde{\mathcal{O}}(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{\varepsilon^2\rho^2\underline{p}^2(1-\gamma)^4})$	$\widetilde{\mathcal{O}}(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert^2}{\varepsilon^2\rho^2\underline{p}^2(1-\gamma)^4})$

robust MDPs. Although all of the uncertainty sets can be cast into the family of so-called f-divergence uncertainty sets, we find it difficult to analyze their finite-sample performance by a general calculation technique. Thus, we analyze the statistical performance of different settings separately and summarize our results in the following parts. For practitioners, our sample complexity results indicate how much data is enough for learning a near-optimal policy in a robust MDP, thus guiding the data-collection process. Our sample complexity results can also serve as theoretical guarantees for the optimality of the learned policy, that is, with a fixed dataset we may describe the minimum level of optimality for our learned policy. In addition, our asymptotic results allows practitioners to make statistical inference for optimal robust value functions. Here are some take-home messages from our results:

(a) Sample-efficient results can be guaranteed with (s,a)-rectangular or s-rectangular assumptions in robust MDPs (upper bound of finite-sample results);

(b) Robust MDPs may have a lower sample complexity than original MDPs when the size of uncertainty set is large (upper and lower bounds of finite-sample results);

(c) Robust MDPs under s-rectangular assumption require more samples than that with (s,a)-rectangular assumption (upper bound of finite-sample results);

(d) Statistical inference for optimal robust value function is possible (asymptotic results).

Finite-sample results. A key criterion of evaluating the finite-sample performance is the following deviation:

$$\max_{\pi} V_r^{\pi}(\mu) - V_r^{\widehat{\pi}^*}(\mu). \tag{1}$$

In this paper, we use a uniform convergence analysis to control Eqn. (1):

$$\max_{\pi} V_r^{\pi}(\mu) - V_r^{\widehat{\pi}^*}(\mu) \le 2 \sup_{\pi\in\Pi} \left| V_r^{\pi}(\mu) - \widehat{V}_r^{\pi}(\mu) \right|. \tag{2}$$

When the dataset is obtained by a generative model, we present the sample complexity of achieving an ε deviation bound of Eqn. (1) in different settings in Table 1. The overall performance among the different uncertainty sets is nearly the same up to some logarithmic factors in the (s,a)-rectangular assumption, which is about $\widetilde{\mathcal{O}}(\frac{|\mathcal{S}|^2|\mathcal{A}|}{\varepsilon^2\rho^2(1-\gamma)^4})$.[1] Compared to the most related work [85], which provided an exponential large sample complexity of robust MDPs, we break the misconception that robust MDPs are exponentially harder than original MDPs in terms of $1/(1-\gamma)$. We leave the detailed discussion of comparison in the related work section.

We also derive sample complexity results under s-rectangular assumption, whose theoretical properties are never studied before while it is a significant setting in robust MDPs [74].

[1] We use $\widetilde{\mathcal{O}}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to hide polylogarithmic factors and universal constants.

TABLE 2
The sample complexity of achieving ε deviation bound (1) in the offline dataset. Here $\nu_{\min} = \min_{s,a,\nu(s,a)>0} \nu(s,a)$

Uncertainty set	(s,a)-rectangular (Theorem 9.1 [81])	s-rectangular (Theorem 9.2 [81])
L_1	$\widetilde{\mathcal{O}}(\frac{\lvert\mathcal{S}\rvert(2+\rho)^2}{\nu_{\min}\varepsilon^2\rho^2(1-\gamma)^4})$	$\widetilde{\mathcal{O}}(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert(2+\rho)^2}{\nu_{\min}\varepsilon^2\rho^2(1-\gamma)^4})$
χ^2	$\widetilde{\mathcal{O}}(\frac{\lvert\mathcal{S}\rvert(1+\rho)^2}{\nu_{\min}\varepsilon^2(\sqrt{1+\rho}-1)^2(1-\gamma)^4})$	$\widetilde{\mathcal{O}}(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert^2(1+\rho)^2}{\nu_{\min}\varepsilon^2(\sqrt{1+\rho}-1)^2(1-\gamma)^4})$
KL	$\widetilde{\mathcal{O}}(\frac{\lvert\mathcal{S}\rvert}{\nu_{\min}\varepsilon^2\rho^2\underline{p}^2(1-\gamma)^4})$	$\widetilde{\mathcal{O}}(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{\nu_{\min}\varepsilon^2\rho^2\underline{p}^2(1-\gamma)^4})$

Notably, the sample complexity would enlarge when we assume the uncertainty sets satisfy the s-rectangular assumption in Table 1. The main difference is caused by the fact that the optimal robust policy is deterministic [55] in the (s,a)-rectangular setting while stochastic [74] in the s-rectangular setting. Thus, the uniform bound over the class of all possible policies (including stochastic and deterministic policies) could be worse than that over the class of deterministic policies.

We also extend our analysis from estimation by a generative model to estimation by an offline dataset, which is generated by a given behavior occupancy measure. As long as the concentrability assumption given in [8] holds, the result of sample complexity only changes by a factor of the concentrability coefficient, which can be referred to Table 2.

Lastly, we show that the sample complexity lower bounds of robust MDPs are $\widetilde{\Omega}(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert(1-\gamma)}{\varepsilon^2}\min\{\frac{1}{(1-\gamma)^4},\frac{1}{\rho^4}\})$ for the L_1 ball and $\widetilde{\Omega}(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{\varepsilon^2(1-\gamma)^2}\min\{\frac{1}{1-\gamma},\frac{1}{\rho}\})$ for the χ^2 ball, but the lower bound of the KL uncertainty set is still lack of explicit expression. Both the upper and lower bound results imply that the robust MDPs can have a lower sample complexity than original MDPs with a proper size ρ of uncertainty set.

Asymptotic results. Indeed, the finite-sample results only imply that $\widehat{V}_r^*(\mu)$ is $\widetilde{O}_P(1/\sqrt{n})$, where a logarithmic factor of n exists. It is not sufficient to guarantee the convergence rate of $\widehat{V}_r^*(\mu)$ to be $1/\sqrt{n}$. Thus, statistical inference from finite-sample results is inaccurate and we need more precise asymptotic results for more accurate statistical inference. Our another contribution is showing that $\widehat{V}_r^*(\mu)$ is $\sqrt{n}$-consistent and also asymptotically normal, and then we can derive statistical inference from data directly. We believe our asymptotic results are novel and may open a new approach to statistical inference in robust MDPs.

Empirical studies. Finally, we evaluate our theoretical results on simulation experiments. Under the (s,a)-rectangular assumption, we follow the classical algorithm Robust Value Iteration [32]. Under the s-rectangular assumption, which is usually more difficult to solve, Bisection Algorithm [30] is applied to obtain the near-optimal robust value function. In both settings, our empirical results show that the performance of the near-optimal robust value function is highly correlated with the number of generative samples. In a large sample regime, we also find that the empirical coverage rate (also called confidence level) of the robust value function is consistent with our theories. We leave more details in Section 5.

1.2. *Related work*. In this subsection, we summarize prior works on three topics: offline RL, robust MDPs, and distributionally robust optimization (DRO).

Offline RL. Two most fundamental problems in offline RL are Off-Policy Evaluation (OPE) and Off-Policy Learning (OPL). These two problems assume the agent is unable to interact with the environment but only has access to a given explorable dataset. In terms of OPE whose purpose is to estimate the value function with a given policy, there are mainly three

different methods: Direct Method (DM), Importance Sampling (IS) [29, 44, 48, 69], and Doubly Robust (DR) method [18, 20, 33, 38, 70]. Here we only discuss the most related method DM. For DM, the usual treatment is firstly estimating the reward and transition probability from the offline dataset, and then applying the model estimators to solve the empirical MDP to obtain the value function. Mannor et al. [50] analyzed the bias and variance of the value function estimation by applying frequency estimators of models in tabular MDPs. To tackle large-scale MDP problems, Jong and Stone [37], Grünewälder et al. [26] proposed other methods to estimate the model of dynamics. Bertsekas and Tsitsiklis [5], Dann et al. [10], Duan, Jia and Wang [13] then extended the DM method to the setting of value function approximation by different algorithms, including regression methods. It is more challenging to analyze OPL (or Batch RL) than OPE, especially under function approximation settings, because the goal of OPL is to learn the optimal policy from the given dataset. When certain assumptions are made, many works have discussed the necessary and sufficient conditions for efficient OPL and provided sample-efficient algorithms within different function hypothesis classes [8, 14, 41, 42, 54, 73, 77, 82].

Robust MDPs[3]. Robust MDPs are related to DM in offline RL. The usual approach to solving robust MDPs is estimating the reward and transition probabilities firstly, and running dynamic programming algorithms to obtain near-optimal solutions [32, 55]. Different from the conventional MDPs [61], robust MDPs allow transition probability taking values in an uncertainty set [49, 79] and aim to obtain an optimal robust policy that maximizes the worst-case value function. Xu and Mannor [80], Petrik [58], Ghavamzadeh, Petrik and Chow [23] showed that the solutions of robust MDPs are less sensitive to estimation errors. However, the choice of uncertainty sets still matters with the solutions of robust MDPs. Wiesemann, Kuhn and Rustem [74] concluded that with the $(s,a)/s$-rectangular and convex set assumptions, the computation complexity of obtaining near-optimal solutions is polynomial.

If the uncertainty set is nonrectangular, the problem becomes NP-hard [74]. With the $(s,a)/s$-rectangular set assumptions, many works have provided efficient learning algorithms to obtain near-optimal solutions in different uncertainty sets [30–32, 39, 55, 68, 74]. In addition, Goyal and Grand-Clement [25] considered a more general assumption called the r-rectangular when MDPs have a low dimensional linear representation. And Derman and Mannor [12] also proposed an extension of robust MDPs (called distributionally robust MDPs) under the Wasserstein distance. Qi and Liao [62] considered the statistical theory in the average reward MDPs case, where they construct a L_1 divergence uncertainty set in the space of the visitation distributions. And their problem formulation is different from robust MDPs.

There are few works considering the nonasymptotic performances of optimal robust policy as Eqn. (1) states. Si et al. [65] considered the asymptotic and nonasymptotic behaviors of the optimal robust solutions in the bandit case when only the KL divergence is applied in the uncertainty set. Zhou et al. [85] extended the nonasymptotic results of Si et al. [65] to the infinite horizon RL case. More importantly, Zhou et al. [85] gave a sample complexity bound $\widetilde{\mathcal{O}}(\frac{C|\mathcal{S}|^2}{\nu_{\min}\varepsilon^2\rho^2(1-\gamma)^2})$. However, they only considered the settings when the KL divergence is applied in the uncertainty set and the (s,a)-rectangular assumption is made, while we consider the settings of the KL ball and other uncertainty sets under both the (s,a) and s-rectangular assumptions. In addition, the result of Zhou et al. [85] is exponentially dependent on $\frac{1}{1-\gamma}$, which is hidden in an unspecified parameter $C = \exp(\frac{1}{\beta(1-\gamma)})$. Indeed, the results of

[3]During the revising process of this manuscript, we noted one very latest paper [56] appeared online, which only studies the finite-sample results of robust MDPs under the (s,a)-rectangular assumption. Compared with their finite-sample results, our corresponding results keep the same as theirs when the L_1 uncertainty set is applied. However, our results have a better dependence on $(1-\gamma)^{-1}$ and ε in cases of both the χ^2 and KL uncertainty sets, whereas their bound still has an exponential dependence on $(1-\gamma)^{-1}$ when the KL uncertainty set is applied.

Zhou et al. [85] gave readers a misconception that robust MDPs are exponentially hard than original MDPs in terms of $1/(1-\gamma)$. In this paper, we break this misconception and prove that robust MDPs can be sample efficient and have lower sample complexity than original MDPs. It is also worth pointing out that an unknown parameter β is hidden in C, which is an optimal solution for a convex problem and has no explicit expression. In our work, we improve their results to a polynomial and explicit sample complexity bound, which is shown in Tables 1 and 2.

Distributionally Robust Optimization (*DRO*). Handling uncertainty sets in robust MDPs is relevant with Distributionally Robust Optimization (DRO), where the objective function is minimized with a worst-case loss function. The core motivation of DRO is to deal with the distribution shift of data using different uncertainty sets. Bertsimas, Gupta and Kallus [6], Delage and Ye [11] formulated the uncertainty set by moment conditions, while Ben-Tal et al. [4], Duchi, Glynn and Namkoong [16], Duchi and Namkoong [15], Lam [40], Duchi and Namkoong [17] formulated the uncertainty set by f-divergence balls. In addition, Wozabal [75], Blanchet and Murthy [7], Gao and Kleywegt [22], Lee and Raginsky [43] also considered Wasserstein balls, which is more computationally challenging. The most related work with our results is Duchi and Namkoong [17], which considered the asymptotic and nonasymptotic performances of the empirical minimizer on a population level. However, the result of Duchi and Namkoong [17] is mainly built on the supervised learning scenario, while our results are built on robust MDPs. Recently, a line of works [9, 36, 76] has studied the connection between pessimistic RL and DRO.

2. Preliminaries.

Markov decision processes. A discounted Markov decision process is defined by a 5-tuple $(\mathcal{S}, \mathcal{A}, R, P, \gamma)$, where $\mathcal{S}$ is the state space and $\mathcal{A}$ is the action space. In this paper, we assume both $\mathcal{S}$ and $\mathcal{A}$ are finite discrete spaces. The reward function satisfies $R\colon \mathcal{S}\times\mathcal{A}\to[0,1]$, the transition probability satisfies $P\colon \mathcal{S}\times\mathcal{A}\to\Delta(\mathcal{S})$, where $\Delta(\mathcal{X}) = \{P\colon \sum_{x\in\mathcal{X}} P(x) = 1, P(x)\geq 0\}$ is a set containing all probability measures on a given finite space $\mathcal{X}$, and $\gamma\in[0,1)$ is the discount factor. A stationary policy π is defined as $\pi\colon \mathcal{S}\to\Delta(\mathcal{A})$ and the value function of a policy π is defined as $V_P^\pi(s) = \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^\infty \gamma^t R(s_t,a_t)|s_0 = s]$, where $\tau\sim\pi$ stands for the trajectory $\tau = (s_0, a_0, s_1, a_1, \ldots)$ generated according to policy π and transition probability P. Furthermore, if the initial distribution μ is given, the value function is $V_P^\pi(\mu) = \mathbb{E}_{s_0\sim\mu} V_P^\pi(s_0)$. The goal of learning an MDP is to solve the problem $\max_\pi V_P^\pi(s)$ for all $s\in\mathcal{S}$ or $\max_\pi V_P^\pi(\mu)$. We denote the optimal value $V_P^*(s) := \max_\pi V_P^\pi(s)$.

Robust Markov decision processes. A robust approach to solving MDP is considering the worst MDP case. The robust value function is $V_r^\pi(s) = \inf_{P\in\mathcal{P}} V_P^\pi(s)$, where transition probability P is taken in a given uncertainty set $\mathcal{P}$. The goal of learning a robust MDP is to solve the problem $\max_\pi \inf_{P\in\mathcal{P}} V_P^\pi(s)$ for all $s\in\mathcal{S}$ or $\max_\pi \inf_{P\in\mathcal{P}} V_P^\pi(\mu)$. We denote the optimal robust value function as $V_r^*(s) = \max_\pi \inf_{P\in\mathcal{P}} V_P^\pi(s)$.

Assumptions on uncertainty set $\mathcal{P}$. Even though there are various choices of uncertainty set $\mathcal{P}$, the existence of a stationary robust optimal policy w.r.t. a robust MDP is only guaranteed when some conditions of uncertainty set $\mathcal{P}$ are satisfied. Iyengar [32], Nilim and El Ghaoui [55] proposed the (s,a)-rectangular set assumption on uncertainty set $\mathcal{P}$, which is detailed in Assumption 2.1.

ASSUMPTION 2.1 ((s,a)-rectangular). The uncertainty set $\mathcal{P}$ is called an (s,a)-rectangular set if it satisfies

$$\mathcal{P} = \bigtimes_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathcal{P}_{s,a},$$

where $\mathcal{P}_{s,a}\subseteq\Delta(\mathcal{S})$ and "$\times$" represents the Cartesian product.

It is shown that the optimal robust policy is stationary and deterministic[4] under Assumption 2.1. In addition, Epstein and Schneider [19], Wiesemann, Kuhn and Rustem [74] proposed an extensive version s-rectangular set, which is detailed in Assumption 2.2.

ASSUMPTION 2.2 (s-rectangular). The uncertainty set $\mathcal{P}$ is called an s-rectangular set if it satisfies

$$\mathcal{P} = \bigtimes_{s\in\mathcal{S}} \mathcal{P}_s,$$

where $\mathcal{P}_s \subseteq \Delta(\mathcal{S})^{|\mathcal{A}|}$ and $\Delta(\mathcal{S})^{|\mathcal{A}|} := \{(P_a)_{a\in\mathcal{A}} | P_a \in \Delta(\mathcal{S}), \text{ for all } a \in \mathcal{A}\}$.

It is shown that the optimal robust policy is stationary, while the optimal robust policy could be stochastic[5] instead of deterministic under Assumption 2.2. For a more general uncertainty set, Wiesemann, Kuhn and Rustem [74] mentioned that it could be NP-hard to obtain the optimal robust policy, which could also be nonstationary and stochastic.

Examples of uncertainty set. Currently, the most frequently used uncertainty sets can all be categorized to the f-divergence set as Examples 2.1 and 2.2 state, where P^* is the center transition probability and ρ determines the size of sets. Iyengar [32] used the L_1 uncertainty set when setting $f(t) = |t-1|$. And Nilim and El Ghaoui [55] used the KL uncertainty set when setting $f(t) = t\log t$. In DRO, Duchi and Namkoong [17] used a more general form of $f(t) \propto t^k$ where $k > 1$, while we only consider $k = 2$ in this paper. As we focus on the statistical performances of robust MDPs, we use $\mathcal{P}_{s,a}(\rho)$, $\mathcal{P}_s(\rho)$ and $\mathcal{P}$ to represent the uncertainty sets when true transition probability P^* is applied. And we use $\widehat{\mathcal{P}}_{s,a}(\rho)$, $\widehat{\mathcal{P}}_s(\rho)$ and $\widehat{\mathcal{P}}$ to represent the uncertainty sets when estimated transition probability $\widehat{P}$ is applied.

EXAMPLE 2.1 (f-divergence under the (s,a)-rectangular assumption). For each (s,a) pair, we denote the center probability by $P^*(\cdot|s,a)$ and the size of the set by $\rho > 0$. The f-divergence (s,a)-rectangular set is defined by

$$\mathcal{P}_{s,a}(\rho) = \left\{ P \in \Delta(\mathcal{S}) | P \ll P^*(\cdot|s,a)^{6}, \sum_{s'\in\mathcal{S}} f\left(\frac{P(s')}{P^*(s'|s,a)}\right) P^*(s'|s,a) \le \rho \right\}.$$

EXAMPLE 2.2 (f-divergence under the s-rectangular assumption). For each $s \in \mathcal{S}$, we denote the center probability by $P^*(\cdot|s,a)$ and the size of the set by $\rho > 0$. The f-divergence s-rectangular set is defined by

$$\mathcal{P}_s(\rho) = \left\{ P \in \Delta(\mathcal{S})^{|\mathcal{A}|} | P(\cdot|a) \ll P^*(\cdot|s,a), \sum_{s'\in\mathcal{S},a\in\mathcal{A}} f\left(\frac{P(s'|a)}{P^*(s'|s,a)}\right) P^*(s'|s,a) \le |\mathcal{A}|\rho \right\}.$$

Connection with nonrobust MDPs. In our settings (Examples 2.1 and 2.2), the parameter ρ controls the difference between robust value function V_r^π and nonrobust value function V^π. Intuitively, we would expect a small difference for a small ρ, which is quantified by the following theorem.

[4] A deterministic policy stands for $\pi(a|s) \in \{0,1\}$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$.

[5] A stochastic policy stands for $\pi(a|s) \in [0,1]$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$.

[6] For any two probability measures P, Q supporting on a finite set $\mathcal{X}$, $P \ll Q$ stands for P is absolutely continuous w.r.t. Q, which means for any $x \in \mathcal{X}$, $Q(x) = 0$ implies $P(x) = 0$.

THEOREM 2.1. *If there exists a monotonically increasing and concave function* $h(t): \mathbb{R}_+ \to \mathbb{R}_+$ *such that for any probability distributions* $P, Q \in \Delta(\mathcal{S})$ *with* $P \ll Q$:

$$\sum_{s\in\mathcal{S}} |P(s) - Q(s)| \le h\Big(\sum_{s\in\mathcal{S}} f\Big(\frac{P(s)}{Q(s)}\Big) Q(s)\Big), \tag{3}$$

then for any fixed policy π, *we have*

$$\|V_r^\pi - V_{P^*}^\pi\|_\infty \le \begin{cases} \dfrac{\gamma h(\rho)}{(1-\gamma)^2} & \text{if Example 2.1 is applied,} \\ \dfrac{\gamma|\mathcal{A}|h(\rho)}{(1-\gamma)^2} & \text{if Example 2.2 is applied.} \end{cases}$$

Specifically, if we use $f(t) = |t-1|, (t-1)^2$, *and* $t\log t$, *respectively, then* $h(t) = t, \sqrt{t}$, *and* $\sqrt{2t}$, *respectively.*[7]

Performance gap of robust MDPs. We usually do not have access to the true transition probability P^* but an unbiased estimated transition probability $\widehat{P}$ can be obtained from a dataset. The empirical optimal robust policy is given by $\widehat{\pi}^* = \operatorname{argmax}_\pi \widehat{V}_r^\pi(\mu)$, where $\widehat{V}_r^\pi(\mu) = \inf_{P\in\widehat{\mathcal{P}}} V_P^\pi(\mu)$. To examine the performance of empirical solution $\widehat{\pi}^*$, we evaluate it by the following performance gap:

$$\max_\pi V_r^\pi(\mu) - V_r^{\widehat{\pi}^*}(\mu). \tag{4}$$

Following the uniform convergence argument in statistical learning theory [28, 53], we can bound this gap by a uniform excess risk [53] as Lemma 2.1 states. For any fixed policy π, we note that $\widehat{V}_r^\pi$ is a fixed point of robust Bellman operator $\widehat{\mathcal{T}}_r^\pi$, which is similar to the nonrobust case [61]. Thus, we can further bound the uniform excess risk as Lemma 2.2 states. Indeed, as long as $\widehat{\mathcal{T}}_r^\pi V$ approximates $\mathcal{T}_r^\pi V$ with enough samples for fixed $V \in \mathcal{V}$ and $\pi \in \Pi$, we can bound the supreme of the uniform excess risks by union bound over Π and $\mathcal{V}$.

LEMMA 2.1. *Denote* $\widehat{\pi}^* = \operatorname{argmax}_{\pi\in\Pi} \widehat{V}_r^\pi(\mu)$, *where* $\widehat{V}_r^\pi(\mu) = \inf_{P\in\widehat{\mathcal{P}}} V_P^\pi(\mu)$ *and* $\widehat{\mathcal{P}}$ *is the uncertainty set with* $\widehat{P}$ *applied. Then the following inequality holds*:

$$0 \le \max_\pi V_r^\pi(\mu) - V_r^{\widehat{\pi}}(\mu) \le 2\sup_{\pi\in\Pi} |V_r^\pi(\mu) - \widehat{V}_r^\pi(\mu)|,$$

where $\Pi = \Delta(\mathcal{A})^{|\mathcal{S}|}$ *contains all probability measures in simplex* $\Delta(\mathcal{A})$ *for each* $s \in \mathcal{S}$.

LEMMA 2.2. *Denoting* $V_r^\pi = (V_r^\pi(s))_{s\in\mathcal{S}}$ *and* $\widehat{V}_r^\pi = (\widehat{V}_r^\pi(s))_{s\in\mathcal{S}}$, *we have*

$$\|V_r^\pi - \widehat{V}_r^\pi\|_\infty \le \frac{1}{1-\gamma}\sup_{V\in\mathcal{V}} \|\mathcal{T}_r^\pi V - \widehat{\mathcal{T}}_r^\pi V\|_\infty,$$

where $\mathcal{T}_r^\pi V = R^\pi + \gamma\inf_{P\in\mathcal{P}} P^\pi V$, $\widehat{\mathcal{T}}_r^\pi V = R^\pi + \gamma\inf_{P\in\widehat{\mathcal{P}}} P^\pi V$ *for any* $V \in \mathcal{V} := [0, \frac{1}{1-\gamma}]^{|\mathcal{S}|}$ *and* $R^\pi(s) := \sum_{a\in\mathcal{A}} R(s,a)\pi(a|s)$, $P^\pi(s'|s) := \sum_{a\in\mathcal{A}} P(s'|s,a)\pi(a|s)$.

REMARK 2.1. In this paper, we consider $\widehat{\mathcal{T}}_r^\pi V = R^\pi + \gamma\inf_{P\in\widehat{\mathcal{P}}} P^\pi V$ with a deterministic reward. If $R(s,a)$ is a bounded random variable for each $(s,a) \in \mathcal{S}\times\mathcal{A}$, we could easily obtain that $\sup_{s,a} |\widehat{R}(s,a) - \mathbb{E}R(s,a)| \le \widetilde{\mathcal{O}}(n^{-1/2})$ with high probability by Hoeffding's inequality, which is much smaller than the statistical error incurred by estimation of the transition probability.

[7] The specific result is obtained by Cauchy–Schwarz inequality and Pinsker's inequality, see Sason and Verdú [63] for details.

Now our goal is to evaluate the supreme of $\|\mathcal{T}_r^\pi V - \widehat{\mathcal{T}}_r^\pi V\|_\infty$ over $\mathcal{V}$ and Π. To do so, we need to estimate the sizes of $\mathcal{V}$ and Π to apply concentration inequalities over the whole sets. Noting that $\mathcal{V}$ is an infinite subset of $\mathbb{R}^{|\mathcal{S}|}$, we apply Lemma 2.3 to discretize the value space $\mathcal{V}$ and bound the performance gap. To discretize the policy set Π, we consider two cases. When the (s,a)-rectangular assumption holds, the optimal robust policy is deterministic, leading to the policy class being finite (i.e., $|\Pi| = |\mathcal{A}|^{|\mathcal{S}|}$). However, when the s-rectangular assumption holds, the optimal policy may be stochastic instead of deterministic, which means the policy class is infinite. Thus, we need Lemma 2.4 to help us control the deviation. We also prove that the covering numbers of $\mathcal{V}$ and Π are bounded as Lemma 2.5 states, which can be used to bound the supreme value over $\mathcal{V}_\varepsilon$ and Π_ε.

LEMMA 2.3. *Let $\mathcal{V}_\varepsilon := \mathcal{N}(V, \|\cdot\|_\infty, \varepsilon)$ denote the smallest ε-net of $\mathcal{V}$ w.r.t. norm $\|\cdot\|_\infty$, which satisfies $\forall V \in \mathcal{V}$ there exists a $V_0 \in \mathcal{V}_\varepsilon$ such that $\|V - V_0\|_\infty \le \varepsilon$. Then we have*

$$\sup_{V\in\mathcal{V}} \left\|\mathcal{T}_r^\pi V - \widehat{\mathcal{T}}_r^\pi V\right\|_\infty \le 2\gamma\varepsilon + \sup_{V\in\mathcal{V}_\varepsilon} \left\|\mathcal{T}_r^\pi V - \widehat{\mathcal{T}}_r^\pi V\right\|_\infty.$$

LEMMA 2.4. *Let $\Pi_\varepsilon := \mathcal{N}(\Pi, \|\cdot\|_1, \varepsilon)$ denote the smallest ε-net of Π w.r.t. norm $\|\cdot\|_1$, which satisfies $\forall \pi \in \Pi$ there exists a $\pi_0 \in \Pi_\varepsilon$ such that $\|\pi(\cdot|s) - \pi_0(\cdot|s)\|_1 \le \varepsilon$ for all $s \in \mathcal{S}$. Then we have*

$$\sup_{\pi\in\Pi, V\in\mathcal{V}} \left\|\mathcal{T}_r^\pi V - \widehat{\mathcal{T}}_r^\pi V\right\|_\infty \le \frac{2\gamma\varepsilon}{1-\gamma} + \sup_{\pi\in\Pi_\varepsilon, V\in\mathcal{V}} \left\|\mathcal{T}_r^\pi V - \widehat{\mathcal{T}}_r^\pi V\right\|_\infty.$$

LEMMA 2.5. *The cardinalities of $\mathcal{V}_\varepsilon$ in Lemma 2.3 and Π_ε in Lemma 2.4 can be respectively bounded by*

$$|\mathcal{V}_\varepsilon| \le \left(1 + \frac{1}{(1-\gamma)\varepsilon}\right)^{|\mathcal{S}|} \quad \text{and} \quad |\Pi_\varepsilon| \le \left(1 + \frac{4}{\varepsilon}\right)^{|\mathcal{S}||\mathcal{A}|}.$$

REMARK 2.2. We give a high-level idea on the construction of $\mathcal{V}_\varepsilon$ and Π_ε in Lemma 2.5. For $\mathcal{V}_\varepsilon$, we can just divide $[0, 1/(1-\gamma)]$ into a grid consisting of ε-sized subintervals at each dimension $s \in \mathcal{S}$. For Π_ε, we can use L_1 balls in $\mathbb{R}^{|\mathcal{A}|-1}$ with size ε to cover the entire policy space.

3. Nonasymptotic results. In this section, we assume there is access to a generative model such that for any given pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, it is able to return an arbitrary value of next states s' following probability $P^*(\cdot|s,a)$. Thus, according to the generated samples, we can construct the empirical estimation of transition probability P^* by

$$\widehat{P}(s'|s,a) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}\big(X_k^{s,a} = s'\big), \tag{5}$$

where $\{X_k^{s,a}\}_{k=1}^n$ are i.i.d. samples generated from $P^*(\cdot|s,a)$. Thus, $\widehat{P}$ is an unbiased estimator of P^*. With the generative model, our nonasymptotic results are stated in the following theorems. In our proof, as the dual problems differ for different choices of f, it is unlikely to obtain a unified concentration result covering all the three settings (L_1, χ^2, and KL cases). Before presenting theoretical results, the proof sketch is as follows:

- First, for any fixed $\pi \in \Pi$ and $V \in \mathcal{V}$, we calculate the dual forms of $\mathcal{T}_r^\pi V(s)$ and $\widehat{\mathcal{T}}_r^\pi V(s)$ for all $s \in \mathcal{S}$ for the different uncertainty sets.
- Second, we bound the concentration error $\|\mathcal{T}_r^\pi V - \widehat{\mathcal{T}}_r^\pi V\|_\infty$ for fixed $\pi \in \Pi$ and $V \in \mathcal{V}$ from the dual forms.

- Next, as $\|\mathcal{T}_r^\pi V - \widehat{\mathcal{T}}_r^\pi V\|_\infty$ is Lipschitz w.r.t. $V \in \mathcal{V}$ in norm $\|\cdot\|_\infty$, we can derive a union bound over $V \in \mathcal{V}$ by Lemma 2.3.
- Finally, under the (s,a)-rectangular assumption, the optimal robust policy is deterministic. Thus, we can derive a union bound over the deterministic policy class, which is finite and satisfies $|\Pi| = |\mathcal{A}|^{|\mathcal{S}|}$. However, when we consider the s-rectangular assumption, the optimal robust policy may be stochastic, which leads to the policy class Π to be infinitely large. According to Lemma 2.4, we can also derive a union bound over Π by taking an ε-net of Π.

REMARK 3.1. We can also extend our nonasymptotic results in this section to the setting with an offline dataset, which can be referred to in Section 9 of the Supplementary Material [81].

3.1. *Results with the (s,a)-rectangular assumption.* Taking $f(t) = |t-1|$, $f(t) = (t-1)^2$, and $f(t) = t\log t$ in Example 2.1, respectively, we have the following results when the (s,a)-rectangular assumption holds.

THEOREM 3.1. *Under the (s,a)-rectangular assumption, the following results hold*:

(a) *If $f(t) = |t-1|$ in Example 2.1 (L_1 balls), then with probability $1-\delta$:*

$$\max_\pi V_r^\pi(\mu) - V_r^{\widehat{\pi}}(\mu) \le \frac{2(2+\rho)\gamma\sqrt{|\mathcal{S}|}}{\rho(1-\gamma)^2\sqrt{2n}}\left(2 + \sqrt{\log\frac{4|\mathcal{S}||\mathcal{A}|^2[1+2(2+\rho)\sqrt{2n}]^2}{\delta(2+\rho)}}\right).$$

(b) *If $f(t) = (t-1)^2$ in Example 2.1 (χ^2 balls), then with probability $1-\delta$:*

$$\max_\pi V_r^\pi(\mu) - V_r^{\widehat{\pi}}(\mu) \le \frac{2C^2(\rho)\gamma\sqrt{|\mathcal{S}|}}{(C(\rho)-1)(1-\gamma)^2\sqrt{n}}\left(4+\sqrt{2\log\frac{2|\mathcal{S}||\mathcal{A}|^2[1+4C(\rho)\sqrt{n}]^2}{\delta C^2(\rho)}}\right),$$

where $C(\rho) = \sqrt{1+\rho}$.

(c) *If $f(t) = t\log t$ in Example 2.1 (KL balls), then with probability $1-\delta$:*

$$\max_\pi V_r^\pi(\mu) - V_r^{\widehat{\pi}}(\mu) \le \frac{4\gamma\sqrt{|\mathcal{S}|}}{\rho(1-\gamma)^2\underline{p}\sqrt{n}}\left(1 + \sqrt{\log\frac{2|\mathcal{S}|^2|\mathcal{A}|^2[1+\rho\underline{p}\sqrt{n}]}{\delta}}\right),$$

where $\underline{p} = \min_{P^*(s'|s,a)>0} P^*(s'|s,a)$.

As a brief sum-up, in order to achieve an ε performance gap, the number of generated samples should be $n_{\text{tot}} := n \times |\mathcal{S}||\mathcal{A}| = \widetilde{\mathcal{O}}(\frac{|\mathcal{S}|^2|\mathcal{A}|}{\varepsilon^2\rho^2(1-\gamma)^4})$ in all the cases under the (s,a)-rectangular assumption, up to some logarithmic factors.

And notably, in Theorem 3.1(c), an additional factor $1/\underline{p}$ occurs in the upper bound, which seems unavoidable. From a high-level point of view, the core step in Theorem 3.1(c) can be expressed by bounding the deviation $|\log\frac{1}{n}\sum_i X_i - \log\mu|$ with $\mathbb{E}X_i = \mu$, where the factor $1/\mu$ plays a significant role on the sample complexity. Fortunately, compared to Zhou et al. [85], whose finite-sample result in the KL setting is exponentially dependent on $\frac{1}{1-\gamma}$, our result in the KL setting is only polynomially dependent on $\frac{1}{1-\gamma}$.

3.2. *Results with the s-rectangular assumption.* Taking $f(t) = |t-1|$, $f(t) = (t-1)^2$, and $f(t) = t\log t$ in Example 2.2, respectively, we have the following results when the s-rectangular assumption holds.

THEOREM 3.2. *Under the s-rectangular assumption, the following results hold*:

(a) *If* $f(t) = |t-1|$ *in Example* 2.1 (L_1 *balls*), *then with probability* $1-\delta$:

$$\max_{\pi} V_r^{\pi}(\mu) - V_r^{\widehat{\pi}}(\mu) \le \frac{2\gamma(2+\rho)\sqrt{|\mathcal{S}||\mathcal{A}|}}{\rho(1-\gamma)^2\sqrt{2n}}\left(4+\sqrt{\log\frac{2|\mathcal{S}|(1+2\sqrt{2n}(\rho+4))^3}{\delta}}\right).$$

(b) *If* $f(t) = (t-1)^2$ *in Example* 2.1 (χ^2 *balls*), *then with probability* $1-\delta$:

$$\max_{\pi} V_r^{\pi}(\mu) - V_r^{\widehat{\pi}}(\mu) \le \frac{2\gamma C^2(\rho)\sqrt{|\mathcal{S}||\mathcal{A}|^2}}{(C(\rho)-1)(1-\gamma)^2\sqrt{n}}\left(6+\sqrt{2\log\frac{2|\mathcal{S}|(1+8\sqrt{n}C(\rho))^3}{\delta}}\right),$$

where $C(\rho) = \sqrt{1+\rho}$.

(c) *If* $f(t) = t\log t$ *in Example* 2.1 (*KL balls*), *then with probability* $1-\delta$:

$$\max_{\pi} V_r^{\pi}(\mu) - V_r^{\widehat{\pi}}(\mu) \le \frac{4\gamma\sqrt{|\mathcal{S}||\mathcal{A}|}}{\rho\underline{p}(1-\gamma)^2\sqrt{n}}\left(2+\sqrt{2\log\frac{2|\mathcal{S}|^2|\mathcal{A}|(1+4\rho\underline{p}\sqrt{n})}{\delta}}\right),$$

where $\underline{p} = \min_{P^*(s'|s,a)>0} P^*(s'|s,a)$.

Under the s-rectangular assumption, the optimal robust policy can be stochastic. In this case, the policy class Π is infinitely large. By controlling the deviation through Lemma 2.4, there could be an amplification in the statistical error. In the cases of both the L_1 and KL balls, the total sample complexity to achieve an ε performance gap is $n_{\text{tot}} = \widetilde{\mathcal{O}}(\frac{|\mathcal{S}|^2|\mathcal{A}|^2}{\varepsilon^2\rho^2(1-\gamma)^4})$. But in the case of the χ^2 balls, the total sample complexity is $n_{\text{tot}} = \widetilde{\mathcal{O}}(\frac{|\mathcal{S}|^2|\mathcal{A}|^3}{\varepsilon^2\rho^2(1-\gamma)^4})$, which is larger than others and caused by the specific dual solution of $\mathcal{T}_r^{\pi}V$.

3.3. *Discussion on* ρ. All of the results under the (s,a) and s-rectangular assumptions suggest that the sample complexity would be unbounded when $\rho \to 0$. To illustrate this phenomenon, we consider a simple distributionally robust optimization problem:

$$\inf_{Q} \sum_{i=1}^{|\mathcal{X}|} Q_i V_i, \tag{6}$$

$$\text{s.t.} \sum_{i=1}^{|\mathcal{X}|} f\left(\frac{Q_i}{P_i}\right) P_i \le \rho, \tag{7}$$

$$\sum_{i=1}^{|\mathcal{X}|} Q_i = 1,\ Q_i \ge 0, \quad \forall i = 1,\dots,|\mathcal{X}|. \tag{8}$$

Here we assume $P \in \Delta(\mathcal{X})$ and $P_i > 0$ for all i. In addition, f is a convex function such that $f(1) = 0$ and $V_i \in [0, M]$ for all i. We denote the optimal value of the above problem (6) as $g(P,\rho)$. Now if we have an unbiased estimator $\widehat{P}$ of P, we would like to know the absolute error between $g(P,\rho)$ and $g(\widehat{P},\rho)$. However, we cannot apply concentration inequality to $g(\widehat{P},\rho)$ directly as the randomness is hidden in the constraint (7). Fortunately, we can write the dual problem of $g(P,\rho)$ and prove the strong duality [64]. In this case, the randomness is displayed in the dual objective (9), where $f^*(y) = -\inf_{x\ge 0} f(x) - xy$ is the conjugate function of f. We denote the dual objective (9) as $d(P,\rho)$, That is,

$$\sup_{\lambda\ge 0,\beta\in\mathbb{R}} -\sum_{i=1}^{|\mathcal{X}|} \lambda P_i f^*\left(-\frac{V_i+\beta}{\lambda}\right) - \lambda\rho - \beta. \tag{9}$$

To control the error $|d(\widehat{P},\rho)-d(P,\rho)|$, we have to determine the range of dual variables λ and β based on the specific choice of f. Then we can apply concentration inequalities uniformly over the range of dual variables. However, the range of dual variables will enlarge to infinity when ρ goes to zero. In this case, the uniform concentration inequalities will suffer error amplification, which leads to infinite sample complexity. Alternatively, by Theorem 2.1 (setting $\gamma=0$ in this case) and Hoeffding's inequality [72], we can upper bound the primal error by

$$\begin{aligned}|g(\widehat{P},\rho)-g(P,\rho)| &\le |g(\widehat{P},\rho)-g(\widehat{P},0)|+|g(P,\rho)-g(P,0)|+|g(\widehat{P},0)-g(P,0)|\\ &\le \mathcal{O}\big(h(\rho)\big)+|g(\widehat{P},0)-g(P,0)|\\ &= \mathcal{O}\big(h(\rho)\big)+\widetilde{\mathcal{O}}\Big(\frac{1}{\sqrt{n}}\Big).\end{aligned}$$

In this case, when ρ approaches zero, we can apply the nonrobust results instead. When we extend the analysis to the setting of robust MDPs with $\rho\to 0$, we can alternatively upper bound the performance gap by

$$\begin{aligned}\big\|V_r^*-V_r^{\widehat{\pi}}\big\|_\infty &\le 2\sup_{\pi\in\Pi}\big\|V_r^\pi-\widehat{V}_r^\pi\big\|_\infty\\ &\le 2\sup_{\pi\in\Pi}\big\|V_r^\pi-V_{P^*}^\pi\big\|_\infty+2\sup_{\pi\in\Pi}\big\|V_{P^*}^\pi-V_{\widehat{P}}^\pi\big\|_\infty+2\sup_{\pi\in\Pi}\big\|V_{\widehat{P}}^\pi-\widehat{V}_r^\pi\big\|_\infty\\ &\le \mathcal{O}\Big(\frac{h(\rho)}{(1-\gamma)^2}\Big)+2\sup_{\pi\in\Pi}\big\|V_{P^*}^\pi-V_{\widehat{P}}^\pi\big\|_\infty\\ &= \mathcal{O}\Big(\frac{h(\rho)}{(1-\gamma)^2}\Big)+\widetilde{\mathcal{O}}\Big(\sqrt{\frac{|\mathcal{S}|}{(1-\gamma)^4 n}}\Big),\end{aligned}$$

where the first inequality is due to Lemma 2.1, the second inequality holds by error decomposition, the third inequality holds by Theorem 2.1, and the last equality holds by sample complexity of nonrobust MDPs with a generative model [1, 24]. In other words, we should not expect robustness when $\rho\to 0$, which also coincides with the theoretical results of the lower bound in the next part.

3.4. *Lower bound.* To complement our nonasymptotic analysis, here we provide the lower bound results of robust MDPs with a generative model. The MDP we construct in Theorem 3.3 is a classic 2-state MDP with only one action, which is frequently analyzed in [14, 24]. The details can be found in Section 8 of the Supplementary Material [81].

THEOREM 3.3 (Lower bound). *There exists a class of robust MDPs with a f-divergence uncertainty set, such that for every (ε,δ)-correct robust RL algorithm $\mathscr{A}$, the total number of generated samples needs to be at least*

$$\widetilde{\Omega}\Big(\frac{|\mathcal{S}||\mathcal{A}|(g'(p))^2p(1-p)}{\varepsilon^2(1-\gamma g(p))^4}\Big),$$

where $p\in(0,1)$, $g(p)=\inf_{D_f(q\|p)\le\rho} q$ and $D_f(q\|p)=pf(\frac{q}{p})+(1-p)f(\frac{1-q}{1-p})$.

In Theorem 3.3, the parameter p can take arbitrary values in $(0,1)$ while we always set p close to 1. Next, we give the exact lower bounds in the following corollaries when the L_1 and χ^2 uncertainty sets are considered. However, when we consider the KL uncertainty set, there is no explicit form of lower bound by the fact that there is no closed-form expression of $g(p)$ when $f(t)=t\log t$.

COROLLARY 3.1 (Lower bound for the L_1 case). *Given that* $f(t) = |t-1|$ *and* $p = \frac{2\gamma-1}{\gamma}$ *in Theorem 3.3, the lower bound of sample complexity is*

$$\widetilde{\Omega}\Big(\frac{|\mathcal{S}||\mathcal{A}|(1-\gamma)}{\varepsilon^2}\min\Big\{\frac{1}{(1-\gamma)^4},\frac{1}{\rho^4}\Big\}\Big).$$

COROLLARY 3.2 (Lower bound for the χ^2 case). *Given that* $f(t) = (t-1)^2$ *and* $p = \frac{2\gamma-1}{\gamma}$ *in Theorem 3.3, the lower bound of sample complexity is*

$$\widetilde{\Omega}\Big(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^2}\min\Big\{\frac{1}{1-\gamma},\frac{1}{\rho}\Big\}\Big).$$

From Corollaries 3.1 and 3.2, we observe that when $\rho \le (1-\gamma)$, the lower bound is exactly $\widetilde{\Omega}(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^3})$, which coincides with the lower bound of classic MDPs with a generative model [24]. When $\rho > (1-\gamma)$, the lower bounds are $\widetilde{\Omega}(\frac{|\mathcal{S}||\mathcal{A}|(1-\gamma)}{\varepsilon^2\rho^4})$ for the L_1 case and $\widetilde{\Omega}(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2\rho(1-\gamma)^2})$ for the χ^2 case.

It is worth noting that a gap exists between the upper bound and lower bound. It is because we obtain the upper bound via a uniform convergence analysis over the whole value space $\mathcal{V}$ and policy space Π. If we are able to find the local deviation bound near the optimal robust value function V_r^*, the upper bound can be tighter and the gap may also be closed. Unfortunately, we have no additional information of V_r^* except for the robust Bellman equation $V_r^* = \mathcal{T}_r V_r^*$, which is insufficient to perform a precise local analysis. We think it is an important work to close the gap and we leave it to subsequent works.

4. Asymptotic results. From the theoretical results of Section 3, we obtain that the statistical convergence rate of robust MDPs is $\widetilde{O}_p(1/\sqrt{n})$.[8] In this section, we investigate the asymptotic properties of robust MDPs. Specifically, in the context of robust MDPs, we show that the robust value function $\widehat{V}_r^\pi(\mu)$ (given policy π) and the optimal robust value function $\widehat{V}_r^*(\mu)$ are $\sqrt{n}$-consistent and asymptotically normal in both the $(s,a)/s$-rectangular settings. Before presenting our results, we first give a high-level idea about how we prove the empirical robust value function to be asymptotically normal.

- Firstly, for any fixed policy $\pi \in \Pi$, we prove the empirical robust Bellman noise is asymptotically normal with a variance matrix $\Lambda^\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$:

$$\sqrt{n}\big(\widehat{\mathcal{T}}_r^\pi V_r^\pi - \mathcal{T}_r^\pi V_r^\pi\big) \xrightarrow{d} \mathcal{N}(0,\Lambda^\pi).$$

- Noting that $\widehat{V}_r^\pi = \widehat{\mathcal{T}}_r^\pi \widehat{V}_r^\pi$, we prove that there exists a matrix $\widehat{\mathbf{M}}^\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$, which is the derivative of the operator $I - \widehat{\mathcal{T}}_r^\pi$ at the point V_r^π, such that

$$\begin{aligned}\sqrt{n}\big(\widehat{\mathcal{T}}_r^\pi V_r^\pi - \mathcal{T}_r^\pi V_r^\pi\big) &= \sqrt{n}\big(\widehat{\mathcal{T}}_r^\pi V_r^\pi - \widehat{\mathcal{T}}_r^\pi \widehat{V}_r^\pi\big) - \sqrt{n}\big(\mathcal{T}_r^\pi V_r^\pi - \widehat{\mathcal{T}}_r^\pi \widehat{V}_r^\pi\big)\\ &= \sqrt{n}\big(\widehat{\mathcal{T}}_r^\pi V_r^\pi - \widehat{\mathcal{T}}_r^\pi \widehat{V}_r^\pi\big) - \sqrt{n}\big(V_r^\pi - \widehat{V}_r^\pi\big)\\ &= -\widehat{\mathbf{M}}^\pi \cdot \sqrt{n}\big(V_r^\pi - \widehat{V}_r^\pi\big) + o_P\big(\sqrt{n}\|V_r^\pi - \widehat{V}_r^\pi\|\big).\end{aligned}$$

- Because the LHS above is asymptotically normal, we can prove that $\sqrt{n}(V_r^\pi - \widehat{V}_r^\pi) = O_P(1)$. By proving that $\widehat{\mathbf{M}}^\pi$ is consistent to $\mathbf{M}^\pi$, we obtain the final result:

$$\sqrt{n}\big(V_r^\pi - \widehat{V}_r^\pi\big) \xrightarrow{d} \mathcal{N}\big(0,(\mathbf{M}^\pi)^{-1}\Lambda^\pi(\mathbf{M}^\pi)^{-\top}\big).$$

[8] For any two random variable sequences $\{X_n\}_{n\ge1}$ and $\{Y_n\}_{n\ge1}$, $X_n = o_P(Y_n)$ stands for X_n/Y_n converges to zero in probability as n goes infinity, and $X_n = O_P(Y_n)$ stands for X_n/Y_n is bounded in probability. See [71] for more details.

- Also, we can extend the results to the case of optimal policies and leave the discussion of this result in Sections 4.1 and 4.2:

$$\sqrt{n}\Big(\max_{\pi} V_r^{\pi} - \max_{\pi} \widehat{V}_r^{\pi}\Big) \xrightarrow{d} \mathcal{N}\big(0, (\mathbf{M}^{\pi^*})^{-1}\Lambda^{\pi^*}(\mathbf{M}^{\pi^*})^{-\top}\big).$$

4.1. *Results with the (s,a)-rectangular assumption.* We consider the asymptotic behaviors of robust value function under (s,a)-rectangular assumptions. From Section 3, we can deduce that the estimator $\widehat{V}_r^{\pi}(\mu)$ converges to $V_r^{\pi}(\mu)$ almost surely (also converges in probability) for a given policy π, which can be seen in Section 11 of the Supplementary Material [81]. Furthermore, $\widehat{V}_r^{\pi}(\mu)$ is also asymptotically normal with rate $\sqrt{n}$ by the following results.

THEOREM 4.1. *Without loss of generality, we assume $V_r^{\pi}(s_1) < \cdots < V_r^{\pi}(s_{|\mathcal{S}|})$. Under the (s,a)-rectangular assumption, we have that for any fixed policy $\pi \in \Pi$,*

$$\sqrt{n}\big(\widehat{V}_r^{\pi}(\mu) - V_r^{\pi}(\mu)\big) \xrightarrow{d} \mathcal{N}\big(0, \mu^{\top}(\mathbf{M}^{\pi})^{-1}\Lambda^{\pi}(\mathbf{M}^{\pi})^{-\top}\mu\big),$$

where Λ^{π} is the asymptotic variance of empirical robust Bellman error, and $\mathbf{M}^{\pi}$ is the derivative of the operator $I - \mathcal{T}_r^{\pi}$ at the point V_r^{π}. Specifically, $\Lambda^{\pi} = \operatorname{diag}\{\sigma_1^2(\pi), \ldots, \sigma_{|\mathcal{S}|}^2(\pi)\}$ where $\sigma_s^2(\pi) = \gamma^2 \sum_{a\in\mathcal{A}} \pi(a|s)^2 \sigma^2(P^*(\cdot|s,a), V_r^{\pi})$. And:

(a) *If $f(t) = |t-1|$ in Example 2.1 (L_1 balls), then the (i,j)th element of $\mathbf{M}^{\pi}$ is*

$$\mathbf{M}^{\pi}(i,j) = \mathbf{1}\{i=j\} - \gamma \sum_{a\in\mathcal{A}} \pi(a|s_i)\Bigg[P(s_j|s_i,a)\mathbf{1}\big\{j < K\big(P(\cdot|s_i,a)\big)\big\} - \Bigg(\sum_{k<K(P(\cdot|s_i,a))} P(s_k|s_i,a) - \Big(1-\frac{\rho}{2}\Big)\Bigg)\mathbf{1}\big\{j = K\big(P(\cdot|s_i,a)\big)\big\} + \frac{\rho}{2}\mathbf{1}\{j=1\}\Bigg],$$

where $K(P) := \min\{l \in \mathbb{Z}_+ | \sum_{k\le l} P(s_k) > 1-\rho/2\}$ for any $P \in \Delta(\mathcal{S})$.

And $\sigma^2(P^*(\cdot|s,a), V_r^{\pi}) = (b_{s,a}^{\pi})^{\top}\Sigma_{s,a}b_{s,a}^{\pi}$ where

$$\Sigma_{s,a}(i,j) = -P^*(s_i|s,a)P^*(s_j|s,a) + P^*(s_i|s,a)\mathbf{1}\{i=j\},$$
$$b_{s,a}^{\pi}(i) = -\big(\eta^*\big(P^*(\cdot|s,a), V_r^{\pi}\big) - V_r^{\pi}(s_i)\big)_+,$$

and η^* is the dual solution of $\mathcal{T}_r^{\pi}V_r^{\pi}$.

(b) *If $f(t) = (t-1)^2$ in Example 2.1 (χ^2 balls), then the (i,j)th element of $\mathbf{M}^{\pi}$ is*

$$\mathbf{M}^{\pi}(i,j) = \mathbf{1}\{i=j\} - \gamma \sum_{a} \pi(a|s_i) C(\rho) \frac{P^*(s_j|s_i,a)(\eta^*(P^*(\cdot|s_i,a), V_r^{\pi}) - V_r^{\pi}(s_j))_+}{\sqrt{\sum_{\tilde{s}\in\mathcal{S}} P^*(\tilde{s}|s_i,a)(\eta^*(P^*(\cdot|s_i,a), V_r^{\pi}) - V_r^{\pi}(\tilde{s}))_+^2}},$$

where $C(\rho) = \sqrt{1+\rho}$. And $\sigma^2(P^*(\cdot|s,a), V_r^{\pi}) = (b_{s,a}^{\pi})^{\top}\Sigma_{s,a}b_{s,a}^{\pi}$ where

$$\Sigma_{s,a}(i,j) = -P^*(s_i|s,a)P^*(s_j|s,a) + P^*(s_i|s,a)\mathbf{1}\{i=j\},$$
$$b_{s,a}^{\pi}(i) = -C(\rho)\frac{(\eta^*(P^*(\cdot|s,a), V_r^{\pi}) - V_r^{\pi}(s_i))_+^2}{2\sqrt{\sum_{s'\in\mathcal{S}} P^*(s'|s,a)(\eta^*(P^*(\cdot|s,a), V_r^{\pi}) - V_r^{\pi}(s'))_+^2}},$$

and η^* is the dual solution of $\mathcal{T}_r^{\pi}V_r^{\pi}$.

(c) *If* $f(t) = t\log t$ *in Example* 2.1 (*KL balls*), *then the* (i, j)*th element of* $\mathbf{M}^\pi$ *is*

$$\mathbf{M}^\pi(i,j) = \mathbf{1}\{i=j\} - \gamma \sum_a \pi(a|s_i) \frac{P^*(s_j|s_i,a)\exp(-\frac{V_r^\pi(s_j)}{\lambda^*(P^*(\cdot|s_i,a),V_r^\pi)})}{\sum_{\tilde{s}\in\mathcal{S}} P^*(\tilde{s}|s_i,a)\exp(-\frac{V_r^\pi(\tilde{s})}{\lambda^*(P^*(\cdot|s_i,a),V_r^\pi)})}.$$

And $\sigma^2(P^*(\cdot|s,a), V_r^\pi) = (b_{s,a}^\pi)^\top \Sigma_{s,a} b_{s,a}^\pi$ *where*

$$\Sigma_{s,a}(i,j) = -P^*(s_i|s,a)P^*(s_j|s,a) + P^*(s_i|s,a)\mathbf{1}\{i=j\},$$

$$b_{s,a}^\pi(i) = -\frac{\lambda^*(P^*(\cdot|s,a), V_r^\pi)\exp(-\frac{V_r^\pi(s_i)}{\lambda^*(P^*(\cdot|s,a),V_r^\pi)})}{\sum_{s'\in\mathcal{S}} P^*(s'|s,a)\exp(-\frac{V_r^\pi(s')}{\lambda^*(P(\cdot|s,a),V_r^\pi(s'))})},$$

and λ^* *is the dual solution of* $\mathcal{T}_r^\pi V_r^\pi$.

Notably, the asymptotic variance is determined by robust value function $V_r^\pi(\mu)$, P^* and optimal dual variables λ^*, η^* of robust optimization problem $\mathcal{T}_r^\pi V_r^\pi$. To estimate the asymptotic variance, we can substitute these variables with consistent estimators $\widehat{V}_r^\pi$, $\widehat{P}$, and $\widehat{\lambda}^*$, $\widehat{\eta}^*$, where $\widehat{\lambda}^*$ and $\widehat{\eta}^*$ are dual solutions of problem $\widehat{\mathcal{T}}_r^\pi \widehat{V}_r^\pi$. Thus, an asymptotic confidence interval for a given policy π can be given by Slutsky's lemma.

We now give the asymptotic results of $\max_\pi \widehat{V}_r^\pi(\mu)$. Prior to that, we define the robust Q-value function $Q_r^\pi(s,a)$ as

$$Q_r^\pi(s,a) = R(s,a) + \gamma \inf_{P\in\mathcal{P}_{s,a}(\rho)} P^\top V_r^\pi.$$

Under some mild assumptions, we show that the asymptotic normality of $\max_\pi \widehat{V}_r^\pi(\mu)$ still holds in the following corollary.

COROLLARY 4.1. *Assuming* $\min_{s,a_1\neq a_2} |Q_r^*(s,a_1) - Q_r^*(s,a_2)| > 0$, *we have*

$$\sqrt{n}\Big(\max_\pi \widehat{V}_r^\pi(\mu) - \max_\pi V_r^\pi(\mu)\Big) \xrightarrow{d} \mathcal{N}\big(0, \mu^\top(\mathbf{M}^{\pi^*})^{-1}\Lambda^{\pi^*}(\mathbf{M}^{\pi^*})^{-\top}\mu\big),$$

where $\pi^* \in \operatorname{argmax}_\pi V_r^\pi(\mu)$, *where* $\mathbf{M}^{\pi^*}$ *and* Λ^{π^*} *are defined in Theorem* 4.1.

Here we give a high-level idea on why Corollary 4.1 holds. When sample size n is large, we can prove that $\widehat{Q}_r^*(s,a)$ approximates $Q_r^*(s,a)$ for each (s,a) pair. In addition, as the (s,a)-rectangular assumption holds, we also know that $\widehat{\pi}^*(s) = \operatorname{argmax}_a \widehat{Q}_r^*(s,a)$ and $\pi^*(s) = \operatorname{argmax}_a Q_r^*(s,a)$ are both deterministic policies. Thus, by Assumption $\min_{s,a_1\neq a_2} |Q_r^*(s,a_1) - Q_r^*(s,a_2)| > 0$, we conclude that $\widehat{\pi}^* = \pi^*$ when sample size n is large. Thus, we can safely consider $\max_\pi \widehat{V}_r^\pi(\mu) = \widehat{V}_r^{\pi^*}(\mu)$ in an asymptotic regime and obtain Corollary 4.1 by applying $\pi = \pi^*$ in Theorems 4.1(a), 4.1(b) and 4.1(c).

4.2. *Results with the* s*-rectangular assumption*. We extend the asymptotic results from the (s,a)-rectangular assumption to the s-rectangular assumption. Unfortunately, when the L_1 uncertainty set is applied, the asymptotic behavior is not guaranteed by the fact that the Bellman operator $\mathcal{T}_r$ is neither differentiable nor affine w.r.t. V. Instead, the asymptotic normality still holds when either the χ^2 uncertainty set or the KL uncertainty set is applied, which is presented as follows.

THEOREM 4.2. *Without loss of generality, we assume* $V_r^\pi(s_1) < \cdots < V_r^\pi(s_{|\mathcal{S}|})$. *Under the* s*-rectangular assumption, we have that for any fixed policy* $\pi \in \Pi$,

$$\sqrt{n}\big(\widehat{V}_r^\pi(\mu) - V_r^\pi(\mu)\big) \xrightarrow{d} \mathcal{N}\big(0, \mu^\top(\mathbf{M}^\pi)^{-1}\Lambda^\pi(\mathbf{M}^\pi)^{-\top}\mu\big),$$

where Λ^π *is the asymptotic variance of empirical robust Bellman error, and* $\mathbf{M}^\pi$ *is the derivative of the operator* $I - \mathcal{T}_r^\pi$ *at the point* V_r^π.

Specifically, $\Lambda^\pi = \operatorname{diag}\{\sigma_1^2(\pi), \ldots, \sigma_{|\mathcal{S}|}^2(\pi)\}$ *where* $\sigma_s^2(\pi) = \gamma^2\sigma^2(\pi, P^*(\cdot|s,\cdot), V_r^\pi)$. *And*:

(a) *If* $f(t) = (t-1)^2$ *in Example* 2.1 (χ^2 *balls*), *then the* (i,j)*th element of* $\mathbf{M}^\pi$ *is*

$$\mathbf{M}^\pi(i,j) = \mathbf{1}\{i=j\} - \gamma\sqrt{|\mathcal{A}|}C(\rho)\frac{\sum_a \pi(a|s_i)P^*(s_j|s_i,a)(\eta_a^*(P^*(\cdot|s_i,\cdot), V_r^\pi) - \pi(a|s_i)V_r^\pi(s_j))_+}{\sqrt{\sum_{\tilde{s},a} P^*(\tilde{s}|s_i,a)(\eta_a^*(P^*(\cdot|s_i,\cdot), V_r^\pi) - \pi(a|s_i)V_r^\pi(\tilde{s}))_+^2}},$$

where $C(\rho) = \sqrt{1+\rho}$. *And* $\sigma^2(\pi, P^*(\cdot|s,\cdot), V_r^\pi) = (b_s^\pi)^\top \Sigma_s b_s^\pi$ *where for two pairs* (s_i, a_k) *and* (s_j, a_l):

$$\Sigma_s\big((s_i,a_k),(s_j,a_l)\big) = \big(-P^*(s_i|s,a_k)P(s_j|s,a_l) + P(s_i|s,a_k)\mathbf{1}\{s_i=s_j\}\big)\mathbf{1}\{a_k=a_l\},$$

$$b_s^\pi(s_i,a_k) = \frac{-\sqrt{|\mathcal{A}|}C(\rho)(\eta_{a_k}^*(P^*(\cdot|s,\cdot), V_r^\pi) - \pi(a_k|s)V_r^\pi(s_i))_+^2}{2\sqrt{\sum_{s'\in\mathcal{S},a'\in\mathcal{A}} P(s'|s,a')(\eta_{a'}^*(P^*(\cdot|s,\cdot), V_r^\pi) - \pi(a'|s)V_r^\pi(s'))_+^2}},$$

and η^* *is the dual solution of* $\mathcal{T}_r^\pi V_r^\pi$.

(b) *If* $f(t) = t\log t$ *in Example* 2.1 (*KL balls*), *then the* (i,j)*th element of* $\mathbf{M}^\pi$ *is*

$$\mathbf{M}^\pi(i,j) = \mathbf{1}\{i=j\} - \gamma\sum_a \frac{\pi(a|s_i)P^*(s_j|s_i,a)\exp(-\frac{\pi(a|s_i)V_r^\pi(s_j)}{\lambda^*(P^*(\cdot|s_i,\cdot),V_r^\pi)})}{\sum_{\tilde{s}\in\mathcal{S}} P^*(\tilde{s}|s_i,a)\exp(-\frac{\pi(a|s_i)V_r^\pi(\tilde{s})}{\lambda^*(P^*(\cdot|s_i,\cdot),V_r^\pi)})},$$

and $\sigma^2(P^*(\cdot|s,\cdot), V_r^\pi) = (b_s^\pi)^\top \Sigma_s b_s^\pi$ *where for two pairs* (s_i, a_k) *and* (s_j, a_l):

$$\Sigma_s\big((s_i,a_k),(s_j,a_l)\big) = \big(-P^*(s_i|s,a_k)\cdot P(s_j|s,a_l) + P(s_i|s,a_k)\mathbf{1}\{s_i=s_j\}\big)\mathbf{1}\{a_k=a_l\},$$

$$b_s(s_i,a_k) = -\frac{\lambda^*(P^*(\cdot|s,\cdot),V_r^\pi)\exp(-\frac{\pi(a_k|s)V_r^\pi(s_i)}{\lambda^*(P^*(\cdot|s,\cdot),V_r^\pi)})}{\sum_{s'\in\mathcal{S}} P(s'|a_k,s)\exp(-\frac{\pi(a_k|s)V_r^\pi(s')}{\lambda^*(P(\cdot|s,\cdot),V_r^\pi)})},$$

and λ^* *is the dual solution of* $\mathcal{T}_r^\pi V_r^\pi$.

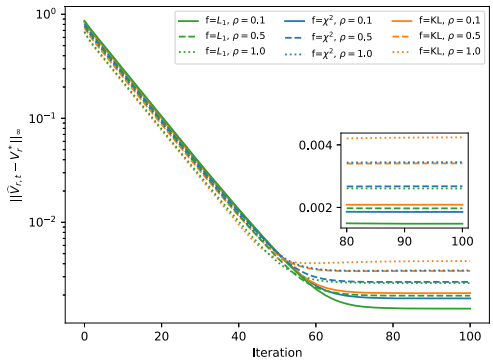
However, different from the (s,a)-rectangular setting, the optimal policies $\widehat{\pi}^* \in \operatorname{argmax}_\pi \widehat{V}_r^\pi$ and $\pi^* \in \operatorname{argmax}_\pi V_r^\pi$ could be stochastic. Thus, we can only obtain $\widehat{\pi}^* \xrightarrow{a.s.} \pi^*$ in the s-rectangular setting (Theorem 4.3) and can not just set $\pi = \pi^*$ in Theorems 4.2(a) and 4.2(b). Fortunately, we could still obtain a result of asymptotic normality of $\widehat{V}_r^*$ in Corollary 4.2, and the details can be found in Section 11 of the Supplementary Material [81].

THEOREM 4.3. *Assuming* $\pi^* \in \operatorname{argmax}_\pi V_r^\pi$ *is unique, we have that* $\widehat{V}_r^* \xrightarrow{a.s.} V_r^*$ *and* $\widehat{\pi}^* \xrightarrow{a.s.} \pi^*$.

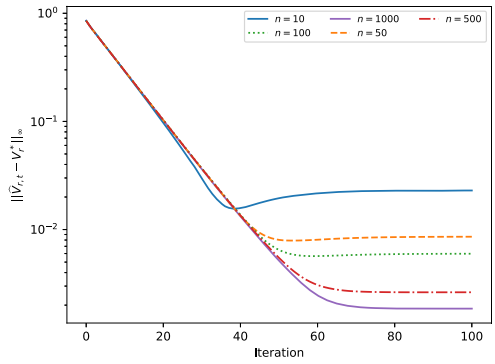
COROLLARY 4.2. *Assuming* π^* *is unique, we have*

$$\sqrt{n}\Big(\max_\pi \widehat{V}_r^\pi(\mu) - \max_\pi V_r^\pi(\mu)\Big) \xrightarrow{d} \mathcal{N}\big(0, \mu^\top(\mathbf{M}^{\pi^*})^{-1}\Lambda^{\pi^*}(\mathbf{M}^{\pi^*})^{-\top}\mu\big),$$

where $\pi^* \in \operatorname{argmax}_\pi V_r^\pi(\mu)$, *where* $\mathbf{M}^{\pi^*}$ *and* Λ^{π^*} *are defined in Theorem* 4.2.



(a) $n = 1000$



(b) $f = \chi^2$, $\rho = 0.1$

FIG. 1. *Convergence results of RVI under* (s, a)*-rectangular settings.* (*a*): *Results of all cases with* $n = 1000$. (*b*): *Results when* $f = \chi^2$ *and* $\rho = 0.1$.

4.3. *Interpretation of asymptotic variance.* The asymptotic variances of robust MDPs under different assumptions all share a similar expression $\mu^\top(\mathbf{M}^\pi)^{-1}\Lambda^\pi(\mathbf{M}^\pi)^{-\top}\mu$, where $\mathbf{M}^\pi$ and Λ^π are determined by the choice of uncertainty sets. The term Λ^π is actually the asymptotic variance of $\sqrt{n}(\widehat{\mathcal{T}}_r^\pi V_r^\pi - \mathcal{T}_r^\pi V_r^\pi)$, which reduces to the variance of empirical Bellman noise [45] in the setting of nonrobust MDPs. Recall that $\mathbf{M}^\pi$ is the derivative of operator $I - \mathcal{T}_r^\pi$ at the point V_r^π, which reduces to the matrix $I - \gamma P^\pi$ in the nonrobust MDPs setting. In other words, the asymptotic variances in robust MDPs share a similar structure with the asymptotic variance in nonrobust MDPs [33, 45]. Besides, the asymptotic results imply the empirical robust value function converges to the true robust value function with a typical rate $\sqrt{n}$. Thus, a direct application of our asymptotic results is to construct a confidence interval for the robust value function, as long as we have a good estimation of asymptotic variances. Indeed, the asymptotic variances have explicit forms in our results (Theorems 4.1 and 4.2). Thus, to construct a confidence interval from the dataset, we can plug empirical estimator $\widehat{P}$ and robust value function $\widehat{V}_r^*$ into the asymptotic variance. We leave the details in the experiments section.

5. Experiments. To evaluate the statistical performance of robust MDPs, we conduct several numerical experiments in this section. We choose randomly generated MDPs as experiment environments. Under the (s, a)-rectangular setting, we run the classic algorithm Robust Value Iteration (RVI) [32] on random MDPs to show that we can obtain a near-optimal value function $\widehat{V}_r^*$ and policy $\widehat{\pi} \in \operatorname{argmax} \widehat{V}_r^\pi$ efficiently. Under the s-rectangular setting, we run the Bisection algorithm, which was proposed by Ho, Petrik and Wiesemann [30]. The details of environments and algorithms are given in Section 12 of the Supplementary Material [81].

5.1. *Convergence guarantees.* We first investigate the convergence performance of RVI on a random MDP, where $|\mathcal{S}| = 20, |\mathcal{A}| = 10$, $\gamma = 0.9$. We leave details of the generation mechanism in Section 12 of the Supplementary Material [81]. For every choice of $f \in \{L_1, \chi^2, \mathrm{KL}\}$, $\rho \in \{0.1, 0.5, 1.0\}$ and $n \in \{10, 50, 100, 500, 1000\}$, we run RVI independently for 1000 times and draw average performances in Figure 1. In Figure 1, the x-axis stands for the number of iteration steps and the y-axis stands for estimation error $\|V_t - V_r^*\|_\infty$, where the V_t come from RVI.

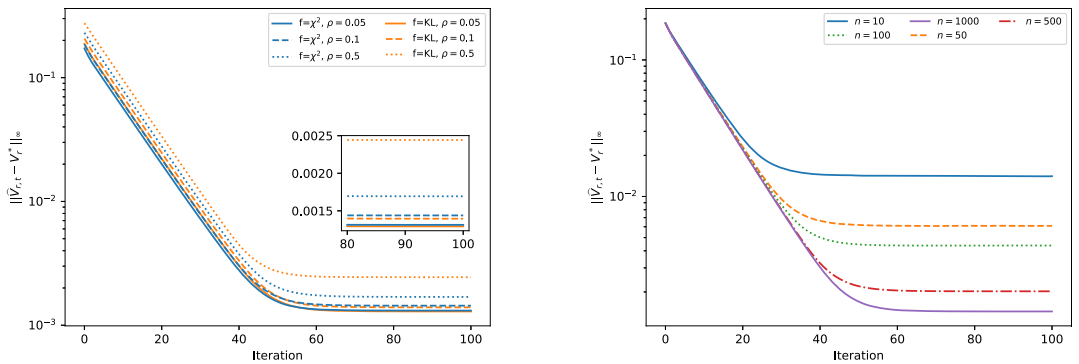
In Figure 1(a), we show the convergence results with all the cases. It can be observed that the convergence rate is linear at the first stage and then becomes stable at a certain error level

for all the settings. Indeed, V_t converges to $\widehat{V}_r^*$ at linear rate and there exist statistical errors between $\widehat{V}_r^*$ and V_r^*. Thus, the first stage in Figure 1(a) is due to linear convergence rate of $\|V_t - \widehat{V}_r^*\|_\infty$ and the second stage is due to statistical error $\|\widehat{V}_r^* - V_r^*\|_\infty$. In Figure 1(b), we set $f = \chi^2$, $\rho = 0.1$ and $n \in \{10, 50, 100, 500, 1000\}$. It is worth noting that the final performance is correlated with the choice of n. In fact, the statistical error $\|\widehat{P}_{s,a} - P_{s,a}\|_1$ is correlated with n. When n is small, it is no wonder that the final performance is bad.

In addition, we run the Bisection algorithm on another random MDP with $|\mathcal{S}| = |\mathcal{A}| = 5$, $\gamma = 0.9$ under the s-rectangular setting. We choose a smaller MDP than the (s, a)-rectangular case by the fact that s-rectangular problems are more difficult to deal with. For every choice of $f \in \{\chi^2, \mathrm{KL}\}$, $\rho \in \{0.05, 0.1, 0.5\}$ and $n \in \{10, 50, 100, 500, 1000\}$, we also run the Bisection algorithm independently for 1000 times and draw average performances in Figure 2. In Figure 2(a), we show the results with all the cases. In comparison with the (s, a)-rectangular setting, the final statistical errors vary less among the different choices of f. In Figure 2(b), we choose $f = \chi^2$, $\rho = 0.1$ and $n \in \{10, 50, 100, 500, 1000\}$. From Figure 2, it is easily observed that the convergence performances are similar with the s-rectangular settings.

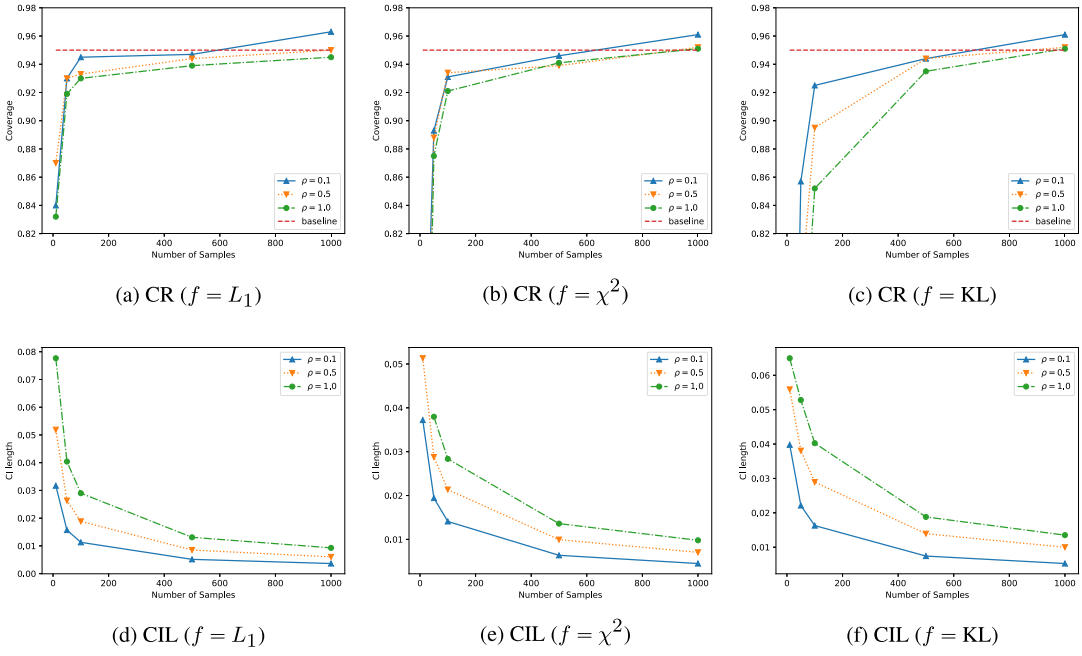
5.2. *Asymptotics*. Next, we follow the theoretical results from Section 4 to make inference on $\widehat{V}_T(\mu)$ empirically under the same settings as in Section 5.1. First of all, based on RVI under the (s, a)-rectangular setting, we estimate $\widehat{\Lambda}$ and $\widehat{\mathbf{M}}^{\pi_T}$ with $\widehat{V}_T$ and obtain $\widehat{\sigma} = \sqrt{\mu^\top (\widehat{\mathbf{M}}^{\pi_T})^{-1} \widehat{\Lambda} (\widehat{\mathbf{M}}^{\pi_T})^{-\top} \mu}$, where Λ and $\mathbf{M}^\pi$ are defined in Section 4. Then we are able to construct a confidence interval $\mathrm{CI}_n(p) = [\widehat{V}_T - z_p \frac{\widehat{\sigma}}{\sqrt{n}}, \widehat{V}_T + z_p \frac{\widehat{\sigma}}{\sqrt{n}}]$, where z_p is the p-quantile of the standard normal distribution $\mathcal{N}(0, 1)$. By the fact that $\widehat{\mathbf{M}}^\pi$ and $\widehat{\Lambda}$ are consistent (refer to the detailed proofs in Section 11 of the Supplementary Material [81]), we can safely say $\lim_{n\to\infty} \mathbb{P}(V_r^*(\mu) \in \mathrm{CI}_n(p)) = 1 - 2(1 - p)$. To evaluate our theory, we test the empirical coverage rate in Table 3 and Figure 3, where we set $p = 0.975$ and $z_p = 1.96$. We observe that the empirical coverage rate approximates the desired true coverage rate and the length of confidence interval decreases as the number of samples increases in all the cases. Interestingly, it seems that the length of confidence interval increases as ρ increases.

Similarly, we also conduct experiments under the s-rectangular setting, where we choose $f \in \{\chi^2, \mathrm{KL}\}$, $\rho \in \{0.05, 0.1, 0.5\}$ and $n \in \{10, 50, 100, 500, 1000\}$, and run the Bisection algorithm on the random MDP ($|\mathcal{S}| = |\mathcal{A}| = 5$) for 1000 times. We also conclude the coverage under the s-rectangular setting in Table 4 and Figure 4, where the results of the empirical coverage meet our expectation.



(a) $f \in \{\chi^2, \mathrm{KL}\}$ (b) $n \in \{10, 50, 100, 500, 1000\}$

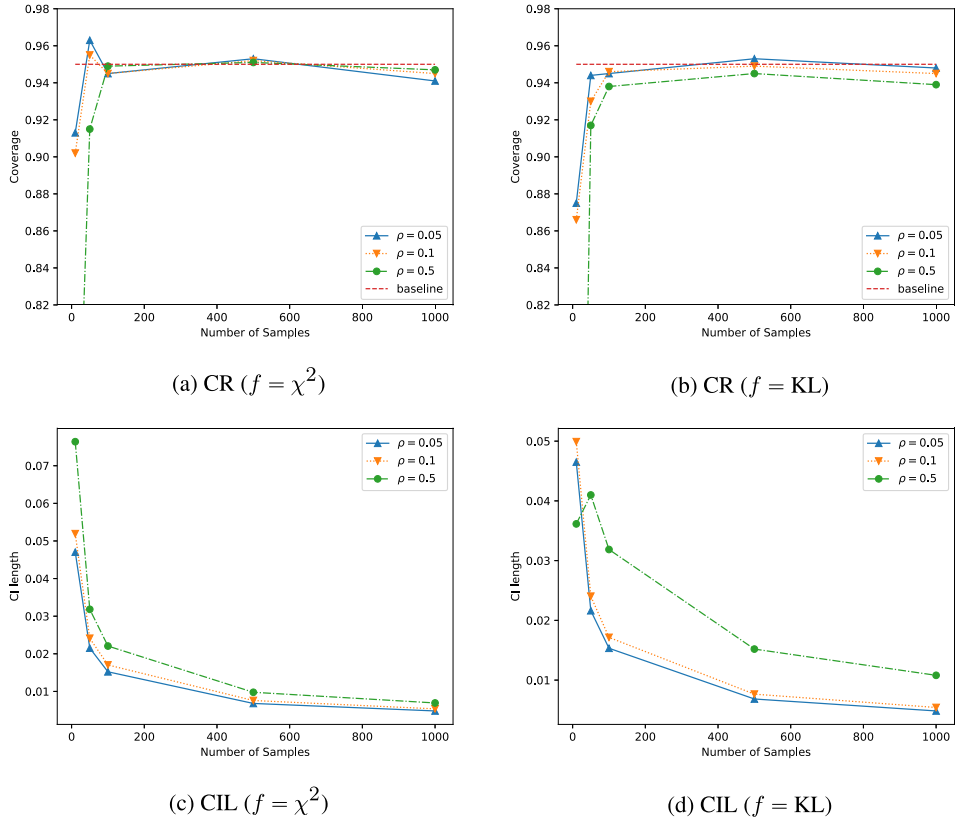
FIG. 2. *Convergence results of Bisection Algorithm under s-rectangular settings.* (*a*): *Results of all cases with $n = 1000$.* (*b*): *Results when $f = \chi^2$ and $\rho = 0.1$.*



(a) CR ($f = L_1$) (b) CR ($f = \chi^2$) (c) CR ($f = \mathrm{KL}$)

(d) CIL ($f = L_1$) (e) CIL ($f = \chi^2$) (f) CIL ($f = \mathrm{KL}$)

FIG. 3. *Coverage rates (CR) and average CI lengths (CIL) under (s, a)-rectangular settings.*



(a) CR ($f = \chi^2$) (b) CR ($f = \mathrm{KL}$)

(c) CIL ($f = \chi^2$) (d) CIL ($f = \mathrm{KL}$)

FIG. 4. *Coverage rates (CR) and average CI lengths (CIL) under s-rectangular settings.*

TABLE 3

Results of coverage rate (CR) and confidence interval length (CIL) under (s, a)-rectangular settings: The standard errors of CR $\widehat{p}$ are computed via $\sqrt{\widehat{p}(1-\widehat{p})/1000} \times 100\%$ and reported inside the parentheses

		$n = 10$			$n = 100$			$n = 1000$		
Items		$\rho = 0.1$	$\rho = 0.5$	$\rho = 1.0$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 1.0$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 1.0$
CR(%)	L_1	84.0 (1.159)	87.0 (1.063)	83.2 (1.182)	94.5 (0.721)	93.3 (0.791)	93.0 (0.801)	96.3 (0.597)	95.0 (0.689)	94.5 (0.721)
	χ^2	64.3 (1.515)	54.9 (1.574)	50.1 (1.581)	93.1 (0.801)	93.4 (0.785)	92.1 (0.853)	96.1 (0.612)	95.2 (0.676)	95.1 (0.683)
	KL	44.8 (1.573)	13.6 (1.084)	5.8 (0.739)	92.5 (0.833)	89.5 (0.969)	85.2 (1.123)	96.1 (0.612)	95.2 (0.676)	95.1 (0.683)
CIL (10^{-2})	L_1	3.170 (0.410)	5.187 (1.590)	7.767 (3.034)	1.129 (0.058)	1.885 (0.188)	2.901 (0.327)	0.365 (0.006)	0.604 (0.019)	0.929 (0.033)
	χ^2	3.722 (0.469)	5.131 (0.851)	6.411 (1.505)	1.409 (0.075)	2.135 (0.163)	2.836 (0.291)	0.450 (0.008)	0.705 (0.019)	0.979 (0.049)
	KL	3.979 (0.566)	5.588 (1.335)	6.496 (2.162)	1.628 (0.102)	2.893 (0.291)	4.025 (0.418)	0.526 (0.011)	0.999 (0.321)	1.351 (0.078)

TABLE 4

Results of coverage rate (CR) and confidence interval length (CIL) under s-rectangular settings: *The standard errors of CR $\widehat{p}$ are computed via $\sqrt{\widehat{p}(1-\widehat{p})/1000} \times 100\%$ and reported inside the parentheses*

		$n = 10$			$n = 100$			$n = 1000$		
Items		$\rho = 0.05$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.05$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.05$	$\rho = 0.1$	$\rho = 0.5$
CR(%)	χ^2	91.3 (0.891)	90.2 (0.940)	68.2 (1.473)	94.5 (0.721)	94.5 (0.721)	94.9 (0.696)	94.1 (0.745)	94.5 (0.721)	94.7 (0.708)
	KL	87.5 (1.046)	86.6 (1.077)	41.9 (1.560)	94.5 (0.721)	94.6 (0.715)	93.8 (0.763)	94.8 (0.702)	94.5 (0.721)	93.9 (0.757)
CIL (10^{-2})	χ^2	4.704 (0.931)	5.193 (1.196)	7.639 (21.657)	1.522 (0.081)	1.703 (0.098)	2.206 (0.336)	0.481 (0.008)	0.536 (0.010)	0.693 (0.031)
	KL	4.650 (1.086)	4.991 (1.545)	3.615 (1.720)	1.533 (0.093)	1.715 (0.140)	3.186 (0.612)	0.484 (0.009)	0.542 (0.014)	1.080 (0.046)

6. Discussion. In this paper, we have studied robust MDPs, which are the foundation of robust RL problems. Our primary concern focuses on the statistical performances of the optimal robust policy and value function obtained from empirical estimation, including finite-sample results and asymptotics based on the most commonly used uncertainty sets: L_1, χ^2, and KL balls. In particular, we have shown that with a polynomial number of samples in the dataset, the performance gap can be controlled well under both the (s, a) and s-rectangular assumptions. Furthermore, we have also shown that the empirical robust optimal value function converges with rate $O_P(1/\sqrt{n})$ and converges to a normal distribution, from which we are able to make inferences from the estimators.

However, some issues still remain open. Firstly, in this paper, the size of the uncertainty set is chosen to be controlled by a positive constant parameter $\rho > 0$. The finite-sample results in Section 3 tell us that a proper choice of ρ can reduce the sample complexity. There are some prior works [12, 59], mentioning that the size of the uncertainty set could also be controlled by a shrinking parameter ρ_n (such as $\rho/\sqrt{n}$), whose statistical properties are still unclear. Thus, understanding the adaptive choice of ρ is a vital topic in future research.

Secondly, in terms of finite-sample results, it is worth noting that there still exists a gap between upper bounds and lower bounds regarding factors $|\mathcal{S}|$, $|\mathcal{A}|$ and $1/(1-\gamma)$. Improving the dependence of these parameters is also a significant research direction.

Finally, in the context of asymptotics, we have proved that $\widehat{V}_r^*(\mu)$ is asymptotically normal with rate $\sqrt{n}$ in both the (s, a) and s-rectangular assumptions. Under the (s, a)-rectangular assumption, the empirical optimal robust policy $\widehat{\pi}^* \in \operatorname{argmax}_\pi \widehat{V}_r^\pi(\mu)$ is exactly the same as $\pi^* \in \operatorname{argmax}_\pi V_r^\pi(\mu)$ when the sample size n is large enough. However, under the s-rectangular assumption, we only know that $\widehat{\pi}^*$ converges to π^* almost surely without a specific convergence rate. According to Van der Vaart [71], we argue that if we could have a more precise estimate of β in the following inequality

$$\mathbb{E} \sup_{d(\pi_1,\pi_2)<\delta} \sqrt{n}\big|\widehat{V}_r^{\pi_1}(\mu) - V_r^{\pi_1}(\mu) - \widehat{V}_r^{\pi_2}(\mu) + V_r^{\pi_2}(\mu)\big| \le C\delta^\beta,$$

the convergence rate of $\widehat{\pi}^*$ then could be determined, and making inference for $\widehat{\pi}^*$ becomes possible. We would leave it to future work.

Acknowledgments. The authors would like to thank the anonymous referees, the Associate Editor and the Editor for their detailed and constructive comments that improved the quality of this paper. The authors would also thank Xiang Li and Dachao Lin for a discussion related to DRO and some inequalities.

Funding. This work has been supported by the National Key Research and Development Project of China (No. 2018AAA0101004).

SUPPLEMENTARY MATERIAL

Supplement to "Toward theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics" (DOI: 10.1214/22-AOS2225SUPP; .pdf). We leave the proofs of theoretical results and details of experiments in the supplementary material.

REFERENCES

[1] AGARWAL, A., KAKADE, S. and YANG, L. F. (2020). Model-based reinforcement learning with a generative model is minimax optimal. In *Proceedings of Thirty Third Conference on Learning Theory* 67–83.

[2] AGARWAL, R., SCHUURMANS, D. and NOROUZI, M. (2020). An optimistic perspective on offline reinforcement learning. In *Proceedings of the* 37*th International Conference on Machine Learning* 104–114.
[3] BEHZADIAN, B., RUSSEL, R. H., PETRIK, M. and HO, C. P. (2021). Optimizing percentile criterion using robust MDPs. In *Proceedings of the* 24*th International Conference on Artificial Intelligence and Statistics* 1009–1017.
[4] BEN-TAL, A., DEN HERTOG, D., DE WAEGENAERE, A., MELENBERG, B. and RENNEN, G. (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Manage. Sci.* **59** 341–357.
[5] BERTSEKAS, D. P. and TSITSIKLIS, J. N. (1995). Neuro-dynamic programming: An overview. In *Proceedings of* 1995 34*th IEEE Conference on Decision and Control* **1** 560–564.
[6] BERTSIMAS, D., GUPTA, V. and KALLUS, N. (2018). Data-driven robust optimization. *Math. Program.* **167** 235–292. MR3755733 https://doi.org/10.1007/s10107-017-1125-8
[7] BLANCHET, J. and MURTHY, K. (2019). Quantifying distributional model risk via optimal transport. *Math. Oper. Res.* **44** 565–600. MR3959085 https://doi.org/10.1287/moor.2018.0936
[8] CHEN, J. and JIANG, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the* 36*th International Conference on Machine Learning* 1042–1051.
[9] DAI, B., NACHUM, O., CHOW, Y., LI, L., SZEPESVÁRI, C. and SCHUURMANS, D. (2020). Coindice: Off-policy confidence interval estimation. *Adv. Neural Inf. Process. Syst.* **33** 9398–9411.
[10] DANN, C., NEUMANN, G. and PETERS, J. (2014). Policy evaluation with temporal differences: A survey and comparison. *J. Mach. Learn. Res.* **15** 809–883. MR3195332
[11] DELAGE, E. and YE, Y. (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.* **58** 595–612. MR2680566 https://doi.org/10.1287/opre.1090.0741
[12] DERMAN, E. and MANNOR, S. (2020). Distributional robustness and regularization in reinforcement learning. ArXiv preprint. Available at arXiv:2003.02894.
[13] DUAN, Y., JIA, Z. and WANG, M. (2020). Minimax-optimal off-policy evaluation with linear function approximation. In *Proceedings of the* 37*th International Conference on Machine Learning* 2701–2709.
[14] DUAN, Y., JIN, C. and LI, Z. (2021). Risk bounds and Rademacher complexity in batch reinforcement learning. In *Proceedings of the* 38*th International Conference on Machine Learning* 2892–2902.
[15] DUCHI, J. and NAMKOONG, H. (2019). Variance-based regularization with convex objectives. *J. Mach. Learn. Res.* **20** Paper No. 68. MR3960922
[16] DUCHI, J. C., GLYNN, P. W. and NAMKOONG, H. (2021). Statistics of robust optimization: A generalized empirical likelihood approach. *Math. Oper. Res.* **46** 946–969. MR4312583 https://doi.org/10.1287/moor.2020.1085
[17] DUCHI, J. C. and NAMKOONG, H. (2021). Learning models with uniform performance via distributionally robust optimization. *Ann. Statist.* **49** 1378–1406. MR4298868 https://doi.org/10.1214/20-aos2004
[18] DUDÍK, M., ERHAN, D., LANGFORD, J. and LI, L. (2014). Doubly robust policy evaluation and optimization. *Statist. Sci.* **29** 485–511. MR3300356 https://doi.org/10.1214/14-STS500
[19] EPSTEIN, L. G. and SCHNEIDER, M. (2003). Recursive multiple-priors. *J. Econom. Theory* **113** 1–31. MR2017864 https://doi.org/10.1016/S0022-0531(03)00097-8
[20] FARAJTABAR, M., CHOW, Y. and GHAVAMZADEH, M. (2018). More robust doubly robust off-policy evaluation. In *Proceedings of the* 35*th International Conference on Machine Learning* 1447–1456.
[21] FUJIMOTO, S., MEGER, D. and PRECUP, D. (2019). Off-policy deep reinforcement learning without exploration. In *Proceedings of the* 36*th International Conference on Machine Learning* 2052–2062.
[22] GAO, R. and KLEYWEGT, A. J. (2016). Distributionally robust stochastic optimization with Wasserstein distance. ArXiv Preprint. Available at arXiv:1604.02199.
[23] GHAVAMZADEH, M., PETRIK, M. and CHOW, Y. (2016). Safe policy improvement by minimizing robust baseline regret. *Adv. Neural Inf. Process. Syst.* **29** 2298–2306.
[24] GHESHLAGHI AZAR, M., MUNOS, R. and KAPPEN, H. J. (2013). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Mach. Learn.* **91** 325–349. MR3064431 https://doi.org/10.1007/s10994-013-5368-1
[25] GOYAL, V. and GRAND-CLEMENT, J. (2022). Robust Markov decision processes: Beyond rectangularity. *Math. Oper. Res.*
[26] GRÜNEWÄLDER, S., LEVER, G., BALDASSARRE, L., PONTIL, M. and GRETTON, A. (2012). Modelling transition dynamics in MDPs with RKHS embeddings. In *Proceedings of the* 29*th International Coference on International Conference on Machine Learning* 1603–1610.
[27] HAARNOJA, T., ZHOU, A., ABBEEL, P. and LEVINE, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the* 35*th International Conference on Machine Learning* 1861–1870.

[28] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294 https://doi.org/10.1007/978-0-387-84858-7

[29] HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. MR1995826 https://doi.org/10.1111/1468-0262.00442

[30] HO, C. P., PETRIK, M. and WIESEMANN, W. (2018). Fast Bellman updates for robust MDPs. In *Proceedings of the* 35*th International Conference on Machine Learning* 1979–1988.

[31] HO, C. P., PETRIK, M. and WIESEMANN, W. (2021). Partial policy iteration for L_1-robust Markov decision processes. *J. Mach. Learn. Res.* **22** Paper No. 275. MR4353054

[32] IYENGAR, G. N. (2005). Robust dynamic programming. *Math. Oper. Res.* **30** 257–280. MR2142033 https://doi.org/10.1287/moor.1040.0129

[33] JIANG, N. and LI, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the* 33*rd International Conference on Machine Learning* 652–661.

[34] JIN, C., ALLEN-ZHU, Z., BUBECK, S. and JORDAN, M. I. (2018). Is Q-learning provably efficient? *Adv. Neural Inf. Process. Syst.* **31**.

[35] JIN, C., YANG, Z., WANG, Z. and JORDAN, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Proceedings of Thirty Third Conference on Learning Theory* 2137–2143.

[36] JIN, Y., YANG, Z. and WANG, Z. (2021). Is pessimism provably efficient for offline rl? In *Proceedings of the* 38*th International Conference on Machine Learning* 5084–5096.

[37] JONG, N. K. and STONE, P. (2007). Model-based function approximation in reinforcement learning. In *Proceedings of the* 6*th International Joint Conference on Autonomous Agents and Multiagent Systems* 1–8.

[38] KALLUS, N. and UEHARA, M. (2020). Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *J. Mach. Learn. Res.* **21** Paper No. 167. MR4209453

[39] KAUFMAN, D. L. and SCHAEFER, A. J. (2013). Robust modified policy iteration. *INFORMS J. Comput.* **25** 396–410. MR3085321 https://doi.org/10.1287/ijoc.1120.0509

[40] LAM, H. (2016). Robust sensitivity analysis for stochastic systems. *Math. Oper. Res.* **41** 1248–1275. MR3544795 https://doi.org/10.1287/moor.2015.0776

[41] LAZARIC, A., GHAVAMZADEH, M. and MUNOS, R. (2012). Finite-sample analysis of least-squares policy iteration. *J. Mach. Learn. Res.* **13** 3041–3074. MR2997720

[42] LE, H., VOLOSHIN, C. and YUE, Y. (2019). Batch policy learning under constraints. In *Proceedings of the* 36*th International Conference on Machine Learning* 3703–3712.

[43] LEE, J. and RAGINSKY, M. (2018). Minimax statistical learning with Wasserstein distances. *Adv. Neural Inf. Process. Syst.* **31**.

[44] LI, L., MUNOS, R. and SZEPESVÁRI, C. (2015). Toward minimax off-policy value estimation. In *Proceedings of the* 18*th International Conference on Artificial Intelligence and Statistics* 608–616.

[45] LI, X., YANG, W., ZHANG, Z. and JORDAN, M. I. (2021). Polyak–Ruppert averaged Q-leaning is statistically efficient. ArXiv Preprint. Available at arXiv:2112.14582.

[46] LILLICRAP, T. P., HUNT, J. J., PRITZEL, A., HEESS, N., EREZ, T., TASSA, Y., SILVER, D. and WIERSTRA, D. (2015). Continuous control with deep reinforcement learning. In *Proceedings of the* 4*th International Conference on Learning Representations*.

[47] LIM, S. H., XU, H. and MANNOR, S. (2013). Reinforcement learning in robust Markov decision processes. *Adv. Neural Inf. Process. Syst.* **26** 701–709.

[48] LIU, Q., LI, L., TANG, Z. and ZHOU, D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Adv. Neural Inf. Process. Syst.* **31**.

[49] MANNOR, S., MEBEL, O. and XU, H. (2012). Lightning does not strike twice: Robust MDPs with coupled uncertainty. In *Proceedings of the* 29*th International Conference on Machine Learning* 451–458.

[50] MANNOR, S., SIMESTER, D., SUN, P. and TSITSIKLIS, J. N. (2004). Bias and variance in value function estimation. In *Proceedings of the* 21*st International Conference on Machine Learning* 72.

[51] MNIH, V., BADIA, A. P., MIRZA, M., GRAVES, A., LILLICRAP, T., HARLEY, T., SILVER, D. and KAVUKCUOGLU, K. (2016). Asynchronous methods for deep reinforcement learning. In *Proceedings of the* 33*rd International Conference on Machine Learning* 1928–1937.

[52] MNIH, V., KAVUKCUOGLU, K., SILVER, D., RUSU, A. A., VENESS, J., BELLEMARE, M. G., GRAVES, A., RIEDMILLER, M., FIDJELAND, A. K. et al. (2015). Human-level control through deep reinforcement learning. *Nature* **518** 529.

[53] MOHRI, M., ROSTAMIZADEH, A. and TALWALKAR, A. (2018). *Foundations of Machine Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3931734

[54] MUNOS, R. and SZEPESVÁRI, C. (2008). Finite-time bounds for fitted value iteration. *J. Mach. Learn. Res.* **9** 815–857. MR2417255

[55] NILIM, A. and EL GHAOUI, L. (2005). Robust control of Markov decision processes with uncertain transition matrices. *Oper. Res.* **53** 780–798. MR2171651 https://doi.org/10.1287/opre.1050.0216

[56] PANAGANTI, K. and KALATHIL, D. (2022). Sample complexity of robust reinforcement learning with a generative model. In *Proceedings of the* 25*th International Conference on Artificial Intelligence and Statistics* 9582–9602.

[57] PENG, X. B., ANDRYCHOWICZ, M., ZAREMBA, W. and ABBEEL, P. (2018). Sim-to-real transfer of robotic control with dynamics randomization. In 2018 *IEEE International Conference on Robotics and Automation* (*ICRA*) 3803–3810.

[58] PETRIK, M. (2012). Approximate dynamic programming by minimizing distributionally robust bounds. In *Proceedings of the* 29*th International Conference on Machine Learning* 1595–1602.

[59] PETRIK, M. and RUSSEL, R. H. (2019). Beyond confidence regions: Tight Bayesian ambiguity sets for robust mdps. *Adv. Neural Inf. Process. Syst.* **32**.

[60] PETRIK, M. and SUBRAMANIAN, D. (2014). RAAM: The benefits of robustness in approximating aggregated MDPs in reinforcement learning. *Adv. Neural Inf. Process. Syst.* **27**.

[61] PUTERMAN, M. L. (1994). *Markov Decision Processes*: *Discrete Stochastic Dynamic Programming*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR1270015

[62] QI, Z. and LIAO, P. (2020). Robust batch policy learning in markov decision processes. ArXiv preprint. Available at arXiv:2011.04185.

[63] SASON, I. and VERDÚ, S. (2016). f-divergence inequalities. *IEEE Trans. Inf. Theory* **62** 5973–6006. MR3565096 https://doi.org/10.1109/TIT.2016.2603151

[64] SHAPIRO, A. (2017). Distributionally robust stochastic programming. *SIAM J. Optim.* **27** 2258–2275. MR3715383 https://doi.org/10.1137/16M1058297

[65] SI, N., ZHANG, F., ZHOU, Z. and BLANCHET, J. (2020). Distributionally robust policy evaluation and learning in offline contextual bandits. In *Proceedings of the* 37*th International Conference on Machine Learning* 8884–8894.

[66] SIDFORD, A., WANG, M., WU, X., YANG, L. F. and YE, Y. (2018). Near-optimal time and sample complexities for solving Markov decision processes with a generative model. In *Advances in Neural Information Processing Systems* **31**.

[67] SILVER, D., HUANG, A., MADDISON, C. J., GUEZ, A., SIFRE, L., VAN DEN DRIESSCHE, G., SCHRITTWIESER, J., ANTONOGLOU, I., PANNEERSHELVAM, V. et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature* **529** 484–489.

[68] SMIRNOVA, E., DOHMATOB, E. and MARY, J. (2019). Distributionally robust reinforcement learning. ArXiv Preprint. Available at arXiv:1902.08708.

[69] SWAMINATHAN, A. and JOACHIMS, T. (2015). The self-normalized estimator for counterfactual learning. *Adv. Neural Inf. Process. Syst.* **28**.

[70] THOMAS, P. and BRUNSKILL, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the* 33*rd International Conference on Machine Learning* 2139–2148.

[71] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

[72] WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics*: *A Non-asymptotic Viewpoint*. *Cambridge Series in Statistical and Probabilistic Mathematics* **48**. Cambridge Univ. Press, Cambridge. MR3967104 https://doi.org/10.1017/9781108627771

[73] WANG, R., FOSTER, D. and KAKADE, S. M. (2020). What are the statistical limits of offline RL with linear function approximation? In *Proceedings of the* 9*th International Conference on Learning Representations*.

[74] WIESEMANN, W., KUHN, D. and RUSTEM, B. (2013). Robust Markov decision processes. *Math. Oper. Res.* **38** 153–183. MR3029483 https://doi.org/10.1287/moor.1120.0566

[75] WOZABAL, D. (2012). A framework for optimization under ambiguity. *Ann. Oper. Res.* **193** 21–47. MR2874755 https://doi.org/10.1007/s10479-010-0812-0

[76] XIAO, C., WU, Y., MEI, J., DAI, B., LATTIMORE, T., LI, L., SZEPESVARI, C. and SCHUURMANS, D. (2021). On the optimality of batch policy optimization algorithms. In *Proceedings of the* 38*th International Conference on Machine Learning* 11362–11371.

[77] XIE, T. and JIANG, N. (2021). Batch value-function approximation with only realizability. In *Proceedings of the* 38*th International Conference on Machine Learning* 11404–11413.

[78] XIE, T., MA, Y. and WANG, Y.-X. (2019). Toward optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Adv. Neural Inf. Process. Syst.* **32**.

[79] XU, H. and MANNOR, S. (2006). The robustness-performance tradeoff in Markov decision processes. *Adv. Neural Inf. Process. Syst.* **19**.

[80] XU, H. and MANNOR, S. (2009). Parametric regret in uncertain Markov decision processes. In *Proceedings of the* 48*h IEEE Conference on Decision and Control* (*CDC*) *Held Jointly with* 2009 28*th Chinese Control Conference* 3606–3613.

[81] YANG, W., ZHANG, L. and ZHANG, Z. (2022). Supplement to "Toward theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics." https://doi.org/10.1214/22-AOS2225SUPP

[82] YIN, M., BAI, Y. and WANG, Y.-X. (2021). Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *Proceedings of the* 24*th International Conference on Artificial Intelligence and Statistics* 1567–1575.

[83] YIN, M. and WANG, Y.-X. (2020). Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *Proceedings of the* 23*rd International Conference on Artificial Intelligence and Statistics* 3948–3958.

[84] ZHAO, W., QUERALTA, J. P. and WESTERLUND, T. (2020). Sim-to-real transfer in deep reinforcement learning for robotics: A survey. In 2020 *IEEE Symposium Series on Computational Intelligence* (*SSCI*) 737–744.

[85] ZHOU, Z., BAI, Q., ZHOU, Z., QIU, L., BLANCHET, J. and GLYNN, P. (2021). Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *Proceedings of the* 24*th International Conference on Artificial Intelligence and Statistics* 3331–3339.