Near-Optimal Distributionally Robust Reinforcement Learning with General L_p Norms

Anonymous Author(s)
Affiliation
Address
email

Abstract

To address the challenges of sim-to-real gap and sample efficiency in reinforcement learning (RL), this work studies distributionally robust Markov decision processes (RMDPs) — optimize the worst-case performance when the deployed environment is within an uncertainty set around some nominal MDP. Despite recent efforts, the sample complexity of RMDPs has remained largely undetermined. While the statistical implications of distributional robustness in RL have been explored in some specific cases, the generalizability of the existing findings remains unclear, especially in comparison to standard RL. Assuming access to a generative model that samples from the nominal MDP, we examine the sample complexity of RMDPs using a class of generalized L_p norms as the 'distance' function for the uncertainty set, under two commonly adopted sa-rectangular and s-rectangular conditions. Our results imply that RMDPs can be more sample-efficient to solve than standard MDPs using generalized L_p norms in both sa- and s-rectangular cases, potentially inspiring more empirical research. We provide a near-optimal upper bound and a matching minimax lower bound for the sa-rectangular scenarios. For s-rectangular cases, we improve the state-of-the-art upper bound and also derive a lower bound using L_∞ norm that verifies the tightness.

1 Introduction

Reinforcement learning (RL) [Sutton, 1988] is a popular paradigm in machine learning, particularly noted for its success in practical applications. The RL framework, usually modeled within the context of a Markov decision process (MDP), focuses on learning effective decision-making strategies based on interactions with an environment. However, the work of Mannor et al. [2004], among others, has highlighted a vulnerability in RL strategies, revealing the sensitivity to estimation errors in the reward and transition probabilities. A specific example of this is when, because of a sim-to-real gap, policies learned in idealized environments catastrophically fail when deployed in settings with slight changes or adversarial perturbations [Klopp et al., 2017, Mahmood et al., 2018].

To address this issue, robust MDPs (RMDPs), proposed by Iyengar [2005] and Nilim and El Ghaoui [2005], have attracted considerable attention. RMDPs are formulated as max-min problems, seeking policies that are resilient to model estimation errors within a specified uncertainty set. Despite the robustness benefits, solving RMDPs is NP-hard for general uncertainty sets [Nilim and El Ghaoui, 2005]. To overcome this challenge, the assumption of rectangularity is often adopted, with uncertainty sets structured as products of independent subsets for each state or state-action pair, denoted as s-rectangular or sa-rectangular assumptions (see Definitions 4 and 5). These assumptions facilitate the use of methods such as robust value iteration and robust policy iteration, preserving many structural properties of MDPs [Ho et al., 2021]. The s-rectangular sets, though less restrictive, pose greater challenges, while the sa-rectangular sets allow for deterministic optimal policies akin

Result type	Reference	Distance	sa-rectangularity		s-rectangularity	
			$0 < \sigma \lesssim 1-\gamma$	$1-\gamma \lesssim \sigma < \sigma_{\max}$	$0 < \tilde{\sigma} \lesssim 1-\gamma$	$1-\gamma \lesssim \tilde{\sigma} < \tilde{\sigma}_{\max}$
Upper bound	Yang et al. [2022a]	TV	$\frac{S^2A(2+\sigma)^2}{\sigma^2(1-\gamma)^4\varepsilon^2}$	$\frac{S^2A(2+\sigma)^2}{\sigma^2(1-\gamma)^4\varepsilon^2}$	$\frac{S^2A^2(2+\tilde{\sigma})^2}{\tilde{\sigma}^2(1-\gamma)^4\varepsilon^2}$	$\frac{S^2A^2(2+\tilde{\sigma})^2}{\tilde{\sigma}^2(1-\gamma)^4\varepsilon^2}$
	Panaganti and Kalathil [2022]	TV	$\frac{S^2A}{(1-\gamma)^4\varepsilon^2}$	$\frac{S^2A}{(1-\gamma)^4\varepsilon^2}$	$\times$	$\times$
	Shi et al. [2023]	TV	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{\sigma(1-\gamma)^2\varepsilon^2}$	$\times$	$\times$
	Clavier et al. [2023]	L_p	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{(1-\gamma)^4\varepsilon^2}$	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{(1-\gamma)^4\varepsilon^2}$
	This paper	L_p	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{\sigma(1-\gamma)^2\varepsilon^2}$	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{(1-\gamma)^2\tilde{\sigma}\min_s\|\pi_s\|_*\varepsilon^2}$
	This paper	General L_p [1]	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{\sigma(1-\gamma)^2\varepsilon^2}$	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{(1-\gamma)^2\tilde{\sigma}C_g\min_s\|\pi_s\|_*\varepsilon^2}$
Lower bound	Yang et al. [2022a]	TV	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA(1-\gamma)}{\sigma^4\varepsilon^2}$	$\times$	$\times$
	Shi et al. [2023]	TV	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{\sigma(1-\gamma)^2\varepsilon^2}$	$\times$	$\times$
	This paper	L_p	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{\sigma(1-\gamma)^2\varepsilon^2}$	$\times$	$\times$
	This paper	L_∞	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{\sigma(1-\gamma)^2\varepsilon^2}$	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{\tilde{\sigma}(1-\gamma)^2\varepsilon^2}$

Table 1: Comparisons with prior results (up to log terms) regarding finding an ε-optimal policy for the distributionally RMDP, where σ is the radius of the uncertainty set and $\sigma_{\max}$ defined in Theorem 1.

to non-robust MDPs [Wiesemann et al., 2013]. Note that, while uncertainty in the reward can be easily handled, dealing with uncertainty in the transition kernel is much more difficult [Kumar et al., 2022, Derman et al., 2021].

The question of sample efficiency is central in RL problems ranging from practice to theory. Although minimax rates are achieved in [Azar et al., 2013b, Li et al., 2023c] in the context of classical MDPs, this goal remains open, in general, in the context of RMDPs. Specifically, there exists prior work studying the sample complexity of distributionally robust RL for a few specific divergences such as total variation (TV), χ^2, KL, and Wasserstein (see a further discussion in Appendix 6) [Yang et al., 2022b, Zhou et al., 2021, Panaganti and Kalathil, 2022], while such results remain unclear for more general classes of L_p norms defined in 1.To this point, to the best of our knowledge, the results of sample complexity that achieve minimax optimality for the full range of uncertainty level are limited to only one case — TV distance [Shi et al., 2023].

In this work, we focus on understanding the sample complexity of RMDPs with a general smooth L_p that will be defined in Def. 1. This generalization is appealing for both practice and theory. In practice, numerous applications are based on optimizations or learning approaches that involve general norms beyond those that have already been studied. Additionally, optimizing norm weighted ambiguity sets for Robust MDPs has been proposed in the context of RMDPs in Russel et al. [2019], which justifies our formulation. Theoretically, prior work has characterized the sample complexity of RMDPs for some specific norms have suggested intriguing insights about the statistical implications of distributional robustness in RL. It is interesting to further understand the statistical cost of robust RL in more general scenarios.One area of focus is the contrast between the sample efficiency of solving distributionally robust RL and solving standard RL. In particular, for the specific case of TV distance, Shi et al. [2023] shows that the sample complexity for solving robust RL is at least the same as and sometimes (when the uncertainty level is relatively large) could be smaller than that of standard RL. This motivates the following open question:

Is distributionally robust RL more sample efficient than standard RL for norms defined in Def. (1) ?

A second question is about the comparisons between the sample complexity of solving s-rectangular RMDPs and that of solving sa-rectangular RMDPs. Note that s-rectangular RMDPs have more complicated optimization formulations with additional variables (uncertainty levels for each action) to optimize. This leads to a richer class of optimal policy candidates—stochastic policies in s-rectangular cases, in contrast to the class of deterministic policies for sa-rectangular cases. In addition, existing sample complexity upper bounds for solving s-rectangular RMDPs are larger than that for solving sa-rectangularity [Yang et al., 2022b] for the investigated cases. This motivates the curious question:

Does solving s-rectangular RMDPs require more samples than solving sa-rectangular RMDPs with general smooth L_p norms defined in Def. 1?

Main contributions. In this paper, we address each of the two questions discussed above. In particular, we provide the first sample complexity analysis for RMDPs with general L_p norms defined in 1 under both the s- and sa-rectangularity conditions. For convenience, we present a detailed comparison between the existing state-of-the-art and our results in Table 1 for quick reference and discuss the contributions and their implications below.

• Considering the first question, we illustrate our results in both sa- and s-rectangular case in Figure 2. In the case of sa-rectangularity, we derive a sample complexity upper bound for RMDPs using general smooth L_p norms (cf. Theorem 1) in the order of $\widetilde{O}\left(\frac{SA}{(1-\gamma)^2\max\{1-\gamma, C_g\sigma\}\varepsilon^2}\right)$. with $C_g > 0$ a positive constant related to the geometry of the norm defined in 1. For classical L_P norms, $C_g \geq 1$ so we can directly relax this constant to 1 to obtain the result in table 1. In addition, we provide a matching minimax lower bound (cf. Theorem 2) that confirms the near-optimality of the upper bound for almost full range of the uncertainty level. Our results match the near-optimal sample complexity derived in Shi et al. [2023] for the specific case using TV distance, while holding for broader cases using general L_p norms. The results rely on a new dual optimization form for sa-rectangular RMDPs and reveal the relationship between the sample complexity and this new dual form — the infinite span seminorm (controlled in Lemma 5), which may be of independent interest.

In the case of s-rectangularity, we provide a sample complexity upper bound for solving RMDPs with general smooth L_p norms in the order of $\widetilde{O}\left(\frac{SA}{(1-\gamma)^2\max\{1-\gamma, C_g \min_s\|\pi_s\|_*\sigma\}\varepsilon^2}\right)$. This result improves the prior art $\widetilde{O}\left(\frac{SA}{(1-\gamma)^4\varepsilon^2}\right)$ in Clavier et al. [2023] for classical L_p when $\tilde{\sigma} \lesssim 1-\gamma$ — by at least a factor of $O\left(\frac{1}{1-\gamma}\right)$. Furthermore, we present a lower bound for a representative case with L_∞ norm, which corroborates the tightness of the upper bound. To the best of our knowledge, this is the first lower bound for solving RMDPs with s-rectangularity.

• Considering the second question, as illustrated in Figure 2, our results highlight that robust RL is at least the same as and sometimes can be more sample-efficient to solve than standard RL for general smooth L_p norms in 1. This insight is of significant practical importance and serves to provide crucial motivation for the use and study of distributionally robustness in RL. Notably, robust RL does not only reduce the vulnerability of RL policy to estimation errors and sim-to-real gaps, but also leads to better data efficiency. In terms of comparing the statistical implications of sa- and s- rectangularity, our results show that solving s-rectangular RMDPs is not harder than solving sa-rectangular RMDPs in terms of sample requirement (See Theorem 3 and Figure 2, Right).

• We highlight the technical contributions as below. For the upper bounds, regarding optimization contribution, we derive new dual optimization problem forms for both $sa-$ and $s-$ rectangular cases(Lemma 3 and 4), which is the foundation of the covering number argument in finite-sample analysis. From a statistical point of view, a new concentration lemma (See Lemma 8 for dual forms and two new lemmas to obtain sample complexity lower than classical RL, controlling the infinite span semi norm of the value function, both for $sa-$ and $s-$ rectangular case are derived (See Lemmas 5 and 6). For the lower bound, the technical contributions are mainly in s-rectangular cases, which involves entire new challenges compared to sa-rectangularity case: the optimal policies can be stochastic and hard to be characterized as a closed form, compared to the deterministic one in sa-rectangular cases. Therefore, we construct new hard instances for s-rectangular cases that is distinct from those used in sa-rectangular cases or standard RL.

2 Problem Formulation: Robust Markov Decision Processes

In this section, we formulate distributionally robust Markov decision processes (RMDPs) in the discounted infinite-horizon setting, introduce the sampling mechanism, and describe our goal.

Standard Markov decision processes (MDPs). A discounted infinite-horizon MDP is represented by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \gamma, P, r)$, where $\mathcal{S} = \{1, \cdots, S\}$ and $\mathcal{A} = \{1, \cdots, A\}$ are the finite state and action spaces, respectively, $\gamma \in [0, 1)$ is the discounted factor, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ denotes the probability transition kernel, and $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the immediate reward function, which is assumed to be deterministic. Moreover, we assume that the reward function is bounded in $(0, 1)$ without loss of generality of the results due to the variance reward invariance. Finally we denote 1_A or 1_S the unitary vector of respectively dimension A or S. Moreover, e_s is the standard unitary vector supported

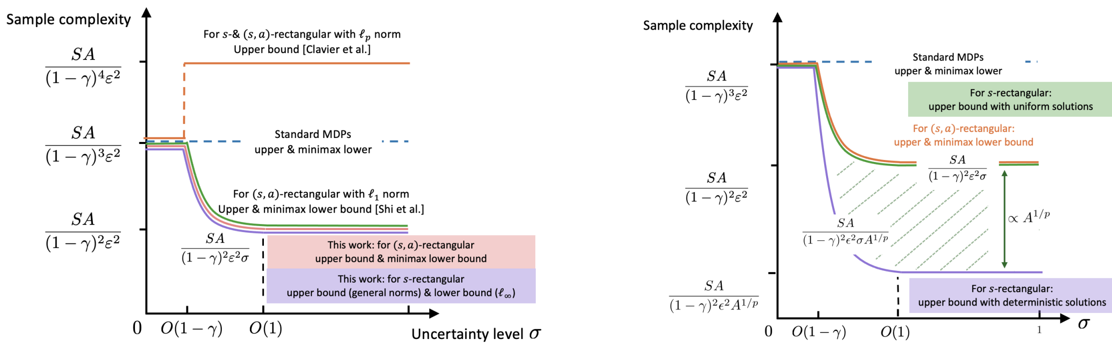

Figure 1: **Left**: Sample complexity results for RMDPs with sa- and s-rectangularity with L_p with comparisons to prior arts [Shi et al., 2023] (for L_1 norm, or called total variation distance) and [Clavier et al., 2023] ; **Right:** The data and instance-dependent sample complexity upper bound of solving s-rectangular dependency RMDPs with L_P norms.

on s. The policy we are looking for is denoted by $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, which specifies the probability of action selection over the action space in any state. Note that if the policy is deterministic in the sa-rectangular case, we overload the notation and refer to $\pi(s)$ as the action selected by the policy π in state s. Finally, to characterize the cumulative reward, the value function $V^{\pi,P}$ for any policy π under the transition kernel P is defined by $\forall s \in \mathcal{S}$

$$V^{\pi,P}(s) := \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t r\big(s_t,a_t\big) \,\Big|\, s_0 = s\right]. \tag{1}$$

The expectation is taken over the randomness of the trajectory $\{s_t, a_t\}_{t=0}^{\infty}$ generated by executing the policy π under the transition kernel P, such that $a_t \sim \pi(\cdot \mid s_t)$ and $s_{t+1} \sim P(\cdot \mid s_t, a_t)$ for all $t \geq 0$. In the same way, the Q function $Q^{\pi,P}$ associated with any policy π under the transition kernel P is defined using expectation taken over the randomness of the trajectory under policy π as

$$Q^{\pi,P}(s,a) := \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t r\big(s_t,a_t\big) \,\Big|\, s_0, a_0 = s, a\right], \tag{2}$$

Distributionally robust MDPs. We consider distributionally robust MDPs (RMDPs) in the discounted infinite-horizon setting, denoted by $\mathcal{M}_{\mathsf{rob}} = \{\mathcal{S}, \mathcal{A}, \gamma, \mathcal{U}^{\sigma}_{\|.\|}(P^0), r\}$, where $\mathcal{S}, \mathcal{A}, \gamma, r$ are the same sets and parameters as in standard MDPs. The main difference compared to standard MDPs is that instead of assuming a fixed transition kernel P, it allows the transition kernel to be arbitrarily chosen from a prescribed uncertainty set $\mathcal{U}^{\sigma}_{\|.\|}(P^0)$ centered around a *nominal* kernel $P^0 : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, where the uncertainty set is specified using some called smooth norm denoted $\|.\|$ defined in of radius $\sigma > 0$ defined in 1.

Definition 1 (General smooth L_p norms and dual norms)**.** *A norm $\|\cdot\|$ is said to be a general smooth L_p norm if*

- *for all $x \in \mathbb{R}^n$, $\|x\| = \|x\|_{p,w} = (\sum_{k=1}^n w_k(|x_k|)^p)^{1/p}$, where $w \in \mathbb{R}^n_+$, is an arbitrary positive vector,*
- *it is twice continuously differentiable Rudin et al. [1964] with the supremum of the Hessian Matrix over the simple $C_S = \sup_{x \in \Delta_s} \left\|\nabla^2 \|x\|\right\|_2$, where $\|\|_2$ here is the spectral norm*

Finally, we denote the dual norm of $\|\cdot\|$ as $\|\cdot\|_*$ s.t. $\|y\|_* = \max_x x^T y : \|x\| \leq 1$. Moreover, for any metric $\|.\|$, we define C_g as $C_g = 1/\min_s \|e_s\|$ where $e_s \in \mathbb{R}^S$ is the standard basis of supported in s.

Note the quantity C_S exists as the Hessian is continuous for C^2 functional and the simplex is a compact set, so by Extreme Value Theorem Rudin et al. [1964], C_S is finite. Moreover, to give an example, considering $L_p, p \geq 2$, norms, C_s is bounded by $S^{1/q}$. (See (151)) This definition is general and includes $L_p, p \geq 2$, all rescaled and weighted norms. Moreover, we could extend our result to a larger set than the one of the norms defined in Def. 1, this is why a complete discussion about the set of norms can be found in Appendix 7. However, it does not include divergences such as KL and χ^2. Not that the case of TV which is not C^2 smooth is treated independently with different arguments in the proof but has the same sample complexity. In particular, given the nominal transition kernel P^0 and some uncertainty level σ, the uncertainty set—with arbitrary smooth norm metric $\| \| : \mathbb{R}^{\mathcal{S}} \times \to \mathbb{R}^+$ in sa rectangular case or from $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ in the s-rectangular case, is specified as $\mathcal{U}^{\sigma}_{\|.\|}(P^0) \coloneqq \otimes_{s,a} \mathcal{U}^{\mathsf{sa},\sigma}_{\|.\|}(P^0_{s,a})$

$$\mathcal{U}^{\mathsf{sa},\sigma}_{\|.\|}(P^0_{s,a}) \coloneqq \left\{ P_{s,a} \in \Delta(\mathcal{S}) : \left\| P_{s,a} - P^0_{s,a} \right\| \leq \sigma \right\}, \tag{3}$$

$$P_{s,a} \coloneqq P(\cdot \mid s,a) \in \mathbb{R}^{1\times S}, P^0_{s,a} \coloneqq P^0(\cdot \mid s,a) \in \mathbb{R}^{1\times S}. \tag{4}$$

where we denote a vector of the transition kernel P or P^0 at state-action pair (s,a). In other words, the uncertainty is imposed in a decoupled manner for each state-action pair, obeying the so-called sa-rectangularity [Zhou et al., 2021, Wiesemann et al., 2013]. More generally, we define s-rectangular MDPs as $\mathcal{U}^{\sigma}_{\|.\|}(P) = \otimes_s \mathcal{U}^{\mathsf{s},\tilde{\sigma}}_{\|.\|}(P_s)$, for the general smooth L_p norm $\|.\|$. The uncertainty is imposed in a decoupled manner for each state pair, and a fixed budget given a state for all action is defined. To get a similar meaning for the radius of the ball between sa-rectangular and s-rectangular assumptions, we need to rescale the radius depending on the norm like in Yang et al. [2022b]. The s- uncertainty set is then defined using the rescaled radius $\tilde{\sigma}$ as

$$\mathcal{U}^{\mathsf{s},\tilde{\sigma}}_{\|.\|}(P_s) \coloneqq \left\{ P'_s \in \Delta(\mathcal{S})^{\mathcal{A}} : \|P'_s - P_s\| \leq \tilde{\sigma} = \sigma \left\| 1_A \right\| \right\}, \tag{5}$$

$$P_s \coloneqq P(\cdot,\cdot \mid s) \in \mathbb{R}^{1\times SA}, \quad P^0_s \coloneqq P^0(\cdot,\cdot \mid s) \in \mathbb{R}^{1\times SA}. \tag{6}$$

where $1_A \in \mathbb{R}^A$ denotes the unitary vector. For the specific case of respectively L_1,L_p and L_∞ norm, $\tilde{\sigma}$ is equal to $|\sigma\mathcal{A}|, \sigma|\mathcal{A}|^{1/p}$ and σ. Note that this scaling allows for a fair comparison between sa- and s-rectangular MDPs. In RMDPs, we are interested in the worst-case performance of a policy π over all the possible transition kernels in the uncertainty set. This is measured by the *robust value function* $V^{\pi,\sigma}$ and the *robust Q-function* $Q^{\pi,\sigma}$ in $\mathcal{M}_{\mathsf{rob}}$, defined respectively as $\forall (s,a) \in \mathcal{S}\times\mathcal{A}$

$$V^{\pi,\sigma}(s) \coloneqq \inf_{P \in \mathcal{U}^{\mathsf{sa},\sigma}_{\|.\|}(P^0)} V^{\pi,P}(s), \quad Q^{\pi,\sigma}(s,a) \coloneqq \inf_{P \in \mathcal{U}^{\mathsf{sa},\sigma}_{\|.\|}(P^0)} Q^{\pi,P}(s,a). \tag{7}$$

Similarly for s-rectangularity, the value function is denoted $V^{\pi,\sigma}_s(s) \coloneqq \inf_{P \in \mathcal{U}^{\mathsf{s},\tilde{\sigma}}_{\|.\|}(P^0)} V^{\pi,P}(s)$.

Optimal robust policy and robust Bellman operator. As a generalization of properties of standard MDPs in the sa-rectangular robust case, it is well-known that there exists at least one deterministic policy that maximizes the robust value function (resp. robust Q-function) simultaneously for all states (resp. state-action pairs) [Iyengar, 2005, Nilim and El Ghaoui, 2005] but not in the s-rectangular case. Therefore, we denote the *optimal robust value function* (resp. *optimal robust Q-function*) as $V^{\star,\sigma}$ (resp. $Q^{\star,\sigma}$), and the optimal robust policy as $\pi^\star$, which satisfy $\forall (s,a) \in \mathcal{S}\times\mathcal{A}$

$$V^{\star,\sigma}(s) \coloneqq V^{\pi^\star,\sigma}(s) = \max_\pi V^{\pi,\sigma}(s), \quad Q^{\star,\sigma}(s,a) \coloneqq Q^{\pi^\star,\sigma}(s,a) = \max_\pi Q^{\pi,\sigma}(s,a). \tag{8a}$$

A key concept in RMDPs is a generalization of Bellman's optimality principle, encapsulated in the following *robust Bellman consistency equation* (resp. *robust Bellman optimality equation*):

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}, \quad Q^{\pi,\sigma}(s,a) = r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{sa},\sigma}_{\|.\|}(P^0_{s,a})} \mathcal{P}V^{\pi,\sigma}, \tag{9a}$$

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A} \quad ,Q^{\star,\sigma}(s,a) = r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{sa},\sigma}_{\|.\|}(P^0_{s,a})} \mathcal{P}V^{\star,\sigma}. \tag{9b}$$

for the sa-rectangular case and same equation replacing $P^0_{s,a}$ by P^0_s and σ by $\tilde{\sigma}$. The robust Bellman operator [Iyengar, 2005, Nilim and El Ghaoui, 2005] is denoted by $\mathcal{T}^\sigma(\cdot) : \mathbb{R}^{SA} \to \mathbb{R}^{SA}$

$$\mathcal{T}^\sigma(Q^\pi)(s,a) \coloneqq r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{sa},\sigma}_{\|.\|}(P^0_{s,a})} \mathcal{P}V, \quad \text{with} \quad V(s) \coloneqq \max_\pi Q^\pi(s,a) \cdot \tag{10}$$

for sa-rectangular MDPs. Given that $Q^{\star,\sigma}$ is the unique-fixed point of $\mathcal{T}^\sigma$ one can recover the optimal robust value function and Q-function using a procedure termed *distributionally robust value iteration* ($DRVI$). Generalizing the standard value iteration, $DRVI$ starts from some given initialization and recursively applies the robust Bellman operator until convergence. As has been shown previously, this procedure converges rapidly due to the γ-contraction property of $\mathcal{T}^\sigma$ with respect to the L_∞ norm [Iyengar, 2005, Nilim and El Ghaoui, 2005].

3 Distributionally Robust Value Iteration

Generative model-based sampling. Following Zhou et al. [2021], Panaganti and Kalathil [2022], we assume access to a generative model or a simulator [Kearns and Singh, 1999], which allows us to collect N independent samples for each state-action pair generated based on the *nominal* kernel P^0: $\forall(s,a) \in \mathcal{S} \times \mathcal{A}$, $s_{i,s,a} \overset{i.i.d}{\sim} P^0(\cdot \mid s,a), \quad i = 1,2,\cdots,N$. The total sample size is, therefore, NSA. We consider a model-based approach tailored to RMDPs, which first constructs an empirical nominal transition kernel based on the collected samples and then applies distributionally robust value iteration (DRVI) to compute an optimal robust policy. As we decouple the statistical estimation error and the optimization error, we exhibit an algorithm that can achieve arbitrary small error ϵ_{opt} in the empirical MDP defined as an empirical nominal transition kernel $\widehat{P}^0 \in \mathbb{R}^{SA\times S}$ that can be constructed on the basis of the empirical frequency of state transitions, i.e. $\forall(s,a) \in \mathcal{S} \times \mathcal{A}$

$$\widehat{P}^0(s' \mid s,a) \coloneqq \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{s_{i,s,a} = s'\}, \tag{11}$$

which leads to an empirical RMDP $\widehat{\mathcal{M}}_{\mathsf{rob}} = \{\mathcal{S},\mathcal{A},\gamma,\mathcal{U}^{\sigma}_{\|\cdot\|}(\widehat{P}^0), r\}$. Analogously, we can define the corresponding robust value function (resp. robust Q-function) of policy π in $\widehat{\mathcal{M}}_{\mathsf{rob}}$ as $\widehat{V}^{\pi,\sigma}$ (resp. $\widehat{Q}^{\pi,\sigma}$) (cf. (8)). In addition, we denote the corresponding *optimal robust policy* as $\widehat{\pi}^{\star}$ and the *optimal robust value function* (resp. *optimal robust Q-function*) as $\widehat{V}^{\star,\sigma}$ (resp. $\widehat{Q}^{\star,\sigma}$) (cf. (9)), which satisfies the robust Bellman optimality equation $\forall(s,a) \in \mathcal{S} \times \mathcal{A}$:

$$\widehat{Q}^{\star,\sigma}(s,a) = r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{sa},\sigma}_{\|\cdot\|}(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}. \tag{12}$$

Equipped with $\widehat{P}^0$, we can define the empirical robust Bellman operator $\widehat{\mathcal{T}}^{\sigma}$ as $\forall(s,a) \in \mathcal{S} \times \mathcal{A}$

$$\widehat{\mathcal{T}}^{\sigma}(Q^{\pi})(s,a) \coloneqq r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{sa},\sigma}_{\|\cdot\|}(\widehat{P}^0_{s,a})} \mathcal{P}V, \tag{13}$$

with $V(s) \coloneqq \max_{\pi} Q^{\pi}(s,a)$. The aim of this work is given the collected samples, to learn the robust optimal policy for the RMDP w.r.t. some prescribed uncertainty set $\mathcal{U}^{\sigma}(P^0)$ around the nominal kernel using as few samples as possible. Specifically, given some target accuracy level $\varepsilon > 0$, the goal is to seek an ε-optimal robust policy $\widehat{\pi}$ obeying

$$\forall s \in \mathcal{S}: \quad V^{\star,\sigma}(s) - V^{\widehat{\pi},\sigma}(s) \leq \varepsilon. \tag{14}$$

$$\widehat{V}^{\widehat{\pi}^{\star},\sigma} - \widehat{V}^{\widehat{\pi},\sigma} \leq \varepsilon_{\mathsf{opt}}. \tag{15}$$

This formulation allows plugging any solver of RMDPs in this bound, for instance, the distributionally robust value iteration (DRVI) algorithm detailed in Appendix 12.

4 Theoretical guarantees

In this section, we present our main results characterizing the sample complexity of solving RMDPs with sa-and s-rectangularity. Additionally, we discuss the implications of our results for the comparisons between standard and robust RL, and for comparisons between sa- versus s-rectangularity.

4.1 sa-rectangular uncertainty set with general smooth norms

To begin, we consider the RMDPs with sa-rectangularity with general norms. We first provide the following sample complexity upper bound for certain oracle planning algorithms, whose proof is postponed to Appendix 9.2. Technically, we derive two new dual forms for RMDPs problems using arbitrary norms in Lemmas 3 and 4 for respectively sa- and s-rectangular RMDPS. In these dual forms, a central quantity denoted $\mathrm{sp}(.)_{*}$, representing the dispersion of the value function, appears and is the dual span semi-norm associated with the considered general L_p norm $\|.\|$ defined in 1 in the initial primal problem. The main challenge in this analysis is to derive a tight upper bound on this quantity in Lemmas (5) and (6), leading to the following sample complexity.

Theorem 1 (Upper bound for sa-rectangularity)**.** *Consider the uncertainty set $\mathcal{U}^{\mathsf{sa},\sigma}_{\|\cdot\|}(\cdot)$ associated with arbitrary smooth norm $\|\cdot\|$ defined in 1. We denote $\sigma_{\max} := \max_{p_1,p_2\in\Delta(\mathcal{S})}\|p_1-p_2\|$ as the accessible maximal uncertainty level. Consider any $\delta\in(0,1)$, discount factor $\gamma\in\left[\frac{1}{4},1\right)$, and uncertainty level $\sigma\in(0,\sigma_{\max}]$. Let $\widehat{\pi}$ be the output policy of some oracle planning algorithm with optimization error $\varepsilon_{\mathsf{opt}}$ introduced in* (15)*. With introduced in 1, one has with probability at least $1-\delta$,*

$$\forall s\in\mathcal{S}:\quad V^{\star,\sigma}(s)-V^{\widehat{\pi},\sigma}(s)\le\varepsilon+\frac{8\varepsilon_{\mathsf{opt}}}{1-\gamma} \tag{16}$$

for any $\varepsilon\in(0,\sqrt{1/\max\{1-\gamma,\sigma C_g\}}]$, as long as the total number of samples obeys

$$NSA\gtrsim\frac{c_1SA}{(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}\varepsilon^2}+\frac{c_2SAC_S\left\|1_S\right\|_*}{(1-\gamma)^2\epsilon} \tag{17}$$

with c_1,c_2,c_3 a universal positive constant. For a sufficiently small level of accuracy $\epsilon\le(\max\{1-\gamma,C_g\sigma\})/(C_S\left\|1_S\right\|)$, the sample complexity is

$$NSA\gtrsim\frac{c_3SA}{(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}\varepsilon^2} \tag{18}$$

Note that this result is also true for TV without the geometric smooth term depending on C_S. Considering L_p norms, $C_g\ge1$ and $C_S\le S^{1/q}$. In Theorem 1, we introduce the following minimax-optimal lower bound to verify the tightness of the above upper bound; a proof is provided in Appendix 10.

Theorem 2 (Lower bound for sa-rectangularity)**.** *Consider the uncertainty set $\mathcal{U}^{\mathsf{sa},\sigma}_{\|\cdot\|}(\cdot)$ associated with arbitrary L_P norm $\|\cdot\|$ defined in* 1*. We denote $\sigma_{\max}:=\max_{p_1,q_1\in\Delta(\mathcal{S})}\|p_1-p_2\|$ as the accessible maximal uncertainty level. Consider any tuple $(S,A,\gamma,\sigma,\varepsilon)$, where $\gamma\in\left[\frac{1}{2},1\right)$, $\sigma\in(0,\sigma_{\max}(1-c_0)]$ with $0<c_0\le\frac{1}{8}$ being any small enough positive constant, and $\varepsilon\in\left(0,\frac{c_0}{256(1-\gamma)}\right]$. We can construct two infinite-horizon RMDPs $\mathcal{M}_0,\mathcal{M}_1$ such that giving a dataset with N independent samples for each state-action pair over the nominal transition kernel (for either $\mathcal{M}_0$ or $\mathcal{M}_1$ respectively), one has*

$$\inf_{\widehat{\pi}}\max_{\mathcal{M}\in\{\mathcal{M}_0,\mathcal{M}_1\}}\left\{\mathbb{P}_{\mathcal{M}}\left(\max_{s\in\mathcal{S}}\left[V^{\star,\sigma}(s)-V^{\widehat{\pi},\sigma}(s)\right]>\varepsilon\right)\right\}\ge\frac{1}{8},$$

where the infimum is taken over all estimators $\widehat{\pi}$, $\mathbb{P}_0$ (resp. $\mathbb{P}_1$) are the probability when the RMDP is $\mathcal{M}_0$ (resp. $\mathcal{M}_1$), as long as, for c_7 is a universal positive constant,

$$NSA\le\frac{c_7SA}{(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}\varepsilon^2}. \tag{19}$$

- **Near minimax-optimal sample complexity with general L_p norms.** Recall that Theorem 1 shows that the sample complexity upper bound of oracle algorithms for RMDPs is in the order of $\widetilde{O}\left(\frac{SA}{(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}\varepsilon^2}\right)$. Combined with the lower bound in Theorem 2, we observe that the above sample complexity is near minimax-optimal, in almost the full range of uncertainty.

- **Solving RMDPs with general L_p norms can be easier than solving standard RL.** Recall that the sample complexity of solving standard RL with a generative model [Agarwal et al., 2020, Li et al., 2024, Azar et al., 2013a] is: $\widetilde{O}\left(\frac{SA}{(1-\gamma)^3\varepsilon^2}\right)$. Comparing this with the sample complexity in (18), it highlights that solving robust MDPs (cf. (18)) using any norm as the divergence function for the uncertainty set is not harder than (and is sometimes easier than) solving standard RL (cf. (4.1)). Specifically, when the uncertainty level is small $\sigma\lesssim1-\gamma$, the sample complexity of solving robust MDPs matches that of standard MDPs. While when the uncertainty level is relatively larger $1-\gamma\lesssim\sigma\le\sigma_{\max}$, the sample complexity of solving robust MDPs is smaller than that of standard MDPs by a factor or $\frac{\sigma}{1-\gamma}$, which goes to $\frac{1}{1-\gamma}$ when $\sigma=O(1)$.

- **Comparisons with prior arts.** In Figure 2, we illustrate the comparisons with two state-of-the-arts [Clavier et al., 2023, Shi et al., 2023] which use some divergence functions belonging to the class of general norms considered in this work. In particular, Shi et al. [2023] achieved the state-of-the-art minimax-optimal sample complexity $\widetilde{O}\left(\frac{SA}{(1-\gamma)^2\max\{1-\gamma,\sigma\}\varepsilon^2}\right)$ for specific L_1 norm (or called total variation distance). In this work, we attain near minimax-optimal sample complexity for any general norm (including L_1) which matches the one in Shi et al. [2023] when narrowing down to L_1 norm. Note that in TV case, $C_g = 1$. This reveals that the finding of robust MDPs can be easier than standard MDPs [Shi et al., 2023] in terms of sample requirement does not only hold for L_1 norm, but for any general norm. In addition, compared to Clavier et al. [2023] which focuses on L_p norms for any $1 \le p \le \infty$: when $1-\gamma \lesssim \sigma \le \sigma_{\max}$, we improve the sample complexity $\widetilde{O}(\frac{SA}{(1-\gamma)^4\varepsilon^2})$ to $\widetilde{O}(\frac{SA}{(1-\gamma)^2\sigma\varepsilon^2})$ by at least a factor of $\frac{1}{1-\gamma}$; otherwise, we match the results in Clavier et al. [2023].

Burn-in Condition, C_g factor and TV case : In Th. 1 and 3 we need a sufficiently small level of accuracy $\epsilon \le (\max\{1-\gamma, C_g\sigma\})/(C_s\|1_S\|)$, to obtain the sample complexity. This type of condition is usual in MDPS analysis Shi et al. [2022] and is equivalent to burn in term. Moreover, the quantity C_S exists (see 1) and for example, considering L_p norms, C_S is bounded by $S^{1/q}$. (See (151)) and the product $C_S\|1_S\|$ is upper bounded by S for L_2 norm. Moreover, note that our theorem for the smooth norm is also true for TV which is not C^2 and has the same complexity as (Shi et al. [2023]. In this case, the burn-in condition is not needed. (See Lemma 9.3.3). Finally, the factor $C_g = 1/\min_s\|e_s\|$ is norm dependent and depends on how big the vector e_{s_0} is in the considered norm. Note for classical L_p this quantity is bigger than 1, which reduces the sample complexity.

4.2 s-rectangular uncertainty set with general norms

To continue, we move on to the case when the uncertainty set is constructed under s-rectangularity smooth norm. The following theorem presents the sample complexity upper bound for learning an ϵ-optimal policy for RMDPs with s-rectangularity. A proof is shown in Appendix 9.2.

Theorem 3 (Upper bound for s-rectangularity)**.** *Consider the uncertainty set $\mathcal{U}^{\mathsf{s},\widetilde{\sigma}}_{\|\cdot\|}(\cdot)$ with s-rectangularity. Consider any discount factor $\gamma \in \left[\frac{1}{4}, 1\right)$, the rescaled uncertainty level $\tilde{\sigma} = \sigma\|1_A\|$, and denote $\tilde{\sigma}_{\max} := \|1_A\|\max_{p_1,p_2\in\Delta(\mathcal{S})}\|p_1-p_2\|$ and $\delta \in (0,1)$. Let $\widehat{\pi}$ be the output policy of an arbitrary optimization algorithm with error $\varepsilon_{\mathsf{opt}}$. , with probability at least $1-\delta$, one has for any $\varepsilon \in (0, \sqrt{1/\max\{1-\gamma, C_g\min_s\|\pi_s\|_*\sigma\}}]$, $\forall s\in\mathcal{S}:\quad V^{\star,\widetilde{\sigma}}(s) - V^{\widehat{\pi},\widetilde{\sigma}}(s) \le \varepsilon + \frac{8\varepsilon_{\mathsf{opt}}}{1-\gamma}$ as long as the total number of samples obeys*

$$NSA \gtrsim \frac{c_4 SA}{(1-\gamma)^2\varepsilon^2}\min\left\{\frac{1}{\max\{1-\gamma, C_g\sigma\}}, \frac{1}{\sigma C_g \min\limits_{s\in\mathcal{S}}\left\{\|\pi^*_s\|_*\|1_A\|, \|\hat{\pi}_s\|_*\|1_A\|\right\}}\right\} + \frac{c_5 SAC_S\|1_S\|_*}{(1-\gamma)^2\epsilon} \tag{20}$$

For a sufficiently small accuracy, $\epsilon \le (\max\{1-\gamma, C_g\tilde{\sigma}\})/(C_s\|1_S\|)$ the sample complexity is

$$NSA \gtrsim \frac{c_6 SA}{(1-\gamma)^2\varepsilon^2}\min\left\{\frac{1}{\max\{1-\gamma, C_g\sigma\}}, \frac{1}{\sigma C_g\min_{s\in\mathcal{S}}\left\{\|\pi^*_s\|_*\|1_A\|, \|\hat{\pi}_s\|_*\|1_A\|\right\}}\right\} \tag{21}$$

where $\hat{\pi}_s \in \Delta_A$ denote the policy of the empirical RMPDs at state s, $\pi^*_s \in \Delta_A$ the optimal policy given s of the true RMPDs, $\|.\|_*$ the dual norm and c_4, c_5, c_6 are universal constant. Note that this result is also true for TV without the term depending on smoothness C_S. In addition, we provide the lower bounds for a representative divergence function — L_∞ norm in the following. Note that for classical L_p, $C_S = S^{1/q}$ and C_g can be lower bounded by 1. A proof is provided in Appendix 11.

Theorem 4 (Lower bound for s-rectangularity)**.** *Consider the uncertainty set $\mathcal{U}^{\mathsf{s},\widetilde{\sigma}}_{L_\infty}(\cdot)$ associated with the L_∞ norm. Consider any tuple $(S, A, \gamma, \sigma, \varepsilon)$ and $0 < c_0 \le \frac{1}{8}$ being any small enough positive constant, where $\gamma \in \left[\frac{1}{2}, 1\right)$, and $\varepsilon \in \left(0, \frac{c_0}{256(1-\gamma)}\right]$. Correspondingly, we denote the accessible maximal uncertainty level for $\mathcal{U}^{\mathsf{s},\widetilde{\sigma}}_{L_\infty}(\cdot)$ as $\sigma^{\infty}_{\max} := \max_{p_1,p_1\in\Delta(\mathcal{S})^A}\|p_1-p_2\|_\infty = 1$. Then we can construct a collection of infinite-horizon RMDPs $\mathcal{M}_{L_\infty}$ defined by the uncertainty set with $\mathcal{U}^{\mathsf{s},\widetilde{\sigma}}_{L_\infty}(\cdot)$ so that for any $\sigma \in (0, \sigma^{\infty}_{\max}(1-c_0)]$, and any dataset with in total N_{all} independent samples for all state-action pairs over the nominal transition kernel (for any RMDP inside $\mathcal{M}_{L_\infty}$), one has*

$$\inf_{\widehat{\pi}}\max_{\mathcal{M}\in\mathcal{M}_{L_\infty}}\left\{\mathbb{P}_{\mathcal{M}}\big(\max_{s\in\mathcal{S}}\left[V^{\star,\sigma}(s) - V^{\widehat{\pi},\sigma}(s)\right] > \varepsilon\big)\right\} \ge \frac{1}{8}, \tag{22}$$

provided that for c_8 is a universal positive constant,

$$N_{\mathsf{all}} \leq \frac{c_8 SA}{(1-\gamma)^2 \max\{1-\gamma,\sigma\}\varepsilon^2}. \tag{23}$$

with $\mathbb{P}_{\mathcal{M}}$ the probability when the RMDP is $\mathcal{M}$, and the infimum is taken over all estimators $\widehat{\pi}$.

Now we can present some implications of Theorem 3 and Theorem 4.

• **Robust MDPs with s-rectangularity are at least as easy as sa-rectangularity.** Theorem 3 shows that the sample complexity of solving RMDPs with s-rectangularity does not exceed the order of $\widetilde{O}\left(\frac{SA}{(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}\varepsilon^2}\right)$. This matches the sample complexity for sa-rectangularity (cf. (18)) and indicates that although s-rectangular RMDPs are of a more complicated formulation, solving s-rectangular RMDPs is at least as easy as solving sa-rectangular RMDPs in terms of the sample complexity. In addition to the worst-case sample complexity upper bound, Theorem 3 also provides a data and instance-dependent sample complexity upper bound for s-rectangular RMDPs (cf. in (20)).Taking the divergence function $\|\cdot\| = L_p$ for instance, the data and instance-dependent sample complexity upper bound is

$$\begin{cases} \widetilde{O}\left(\frac{SA}{(1-\gamma)^2\varepsilon^2}\frac{1}{\max\{1-\gamma,\sigma\}}\right) & \text{if } \widehat{\pi}_s(a\,|\,s) = \pi^*_s(a\,|\,s) = \frac{1}{A}, \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A} \\ \widetilde{O}\left(\frac{SA}{(1-\gamma)^2\varepsilon^2}\frac{1}{\max\{1-\gamma,\sigma A^{1/p}\}}\right) & \text{if } \|\widehat{\pi}_s(\cdot\,|\,s)\|_0 = \|\pi^*_s(\cdot\,|\,s)\|_0 = 1, \quad \forall s \in \mathcal{S}. \end{cases}$$

where $\|.\|_0$ corresponds to the total number of nonzero elements in a vector.The intuition beyond this theorem is that when the policy becomes proportional to uniform, the uncertainty budget of the s-rectangular MDPs is equally spread into all actions, and we retrieve the sa-rectangular case. When the policy becomes deterministic, all the uncertainty budget concentrates on one action. In this case, most of the actions are not robust except one, and the problem is simpler than classical MDP for this only specific action. An illustration of this result can be found in Fig. 2.

• **Comparisons with prior arts.** In Figure 2, we illustrate the comparisons with Clavier et al. [2023] which use L_p norms functions belonging to the class of general norms considered in this work. We do not compare in this section to Yang et al. [2022a] as it is not anymore state-of-the-art with regard to the work of Clavier et al. [2023]. In particular, the latest achieves in the s-rectangular case at sample complexity of $\widetilde{O}\left(\frac{SA}{(1-\gamma)^3\varepsilon^2}\right)$ in the regime where $\tilde{\sigma} \lesssim 1-\gamma$. In this regime, our result is the same but more general but in the regime where $\tilde{\sigma} \gtrsim 1-\gamma$, they achieve sample complexity of $\widetilde{O}\left(\frac{SA}{(1-\gamma)^4\varepsilon^2}\right)$ which is bigger than our result $\widetilde{O}\left(\frac{SA}{(1-\gamma)^2\max\{1-\gamma,\sigma\}\varepsilon^2}\right)$ by a factor at least $\frac{1}{1-\gamma}$.

5 Conclusion

This work refined sample complexity bounds to learn robust Markov decision processes when the uncertainty set is characterized by an general L_p metric, assuming the presence of a generative model. Our findings not only strengthen the current knowledge by improving both the upper and lower bounds, but also highlight that learning s-rectangular MDPs is less challenging in terms of sample complexity compared to classical sa-rectangular MDPs. This work is the first to provide results with a minimax bound, as prior results concerning s-rectangular cases were not minimax optimal. Additionally, we have established the minimax sample complexity for RMDPs using a general L_p norm, demonstrating that it is never larger than that required for learning standard MDPs. Our research identifies potential avenues for future work, such as exploring the characterization of tight sample complexity for RMDPs under a broader family of uncertainty sets, such as those defined by f-divergence. It would be highly desirable for a more unified theoretical foundation, as the distance between probability measures is more natural to define using divergence. Moreover, it would be interesting to focus on the finite-horizon Setting and linear setting, as our current analytical framework opens the door for potential extensions to address finite-horizon RMDPs. Such an extension would contribute to a more comprehensive understanding of tabular cases. Finally, the case of linear MDPs would be interesting to explore.

References

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020.

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

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3): 325–349, 2013b.

Kishan Panaganti Badrinath and Dileep Kalathil. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *International Conference on Machine Learning*, pages 511–520. PMLR, 2021.

Yu Bai, Tengyang Xie, Nan Jiang, and Yu-Xiang Wang. Provably efficient Q-learning with low switching cost. *arXiv preprint arXiv:1905.12849*, 2019.

Carolyn L Beck and Rayadurgam Srikant. Error bounds for constant step-size Q-learning. *Systems & control letters*, 61(12):1203–1208, 2012.

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

Jose Blanchet, Miao Lu, Tong Zhang, and Han Zhong. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *arXiv preprint arXiv:2305.09659*, 2023.

Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. Finite-sample analysis of stochastic approximation using smooth convex envelopes. *arXiv preprint arXiv:2002.00874*, 2020.

Pierre Clavier, Stéphanie Allassonière, and Erwan Le Pennec. Robust reinforcement learning with distributional risk-averse formulation. *arXiv preprint arXiv:2206.06841*, 2022.

Pierre Clavier, Erwan Le Pennec, and Matthieu Geist. Towards minimax optimality of model-based robust reinforcement learning. *arXiv preprint arXiv:2302.05372*, 2023.

Esther Derman and Shie Mannor. Distributional robustness and regularization in reinforcement learning. *arXiv preprint arXiv:2003.02894*, 2020.

Esther Derman, Matthieu Geist, and Shie Mannor. Twice regularized MDPs and the equivalence between robustness and regularization. *Advances in Neural Information Processing Systems*, 34, 2021.

Jing Dong, Jingwei Li, Baoxiang Wang, and Jingzhao Zhang. Online policy optimization for robust MDP. *arXiv preprint arXiv:2209.13841*, 2022.

Kefan Dong, Yuanhao Wang, Xiaoyu Chen, and Liwei Wang. Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. *arXiv preprint arXiv:1901.09311*, 2019.

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

Rui Gao. Finite-sample guarantees for wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *arXiv preprint arXiv:2009.04382*, 2020.

Vineet Goyal and Julien Grand-Clement. Robust markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 2022.

Songyang Han, Sanbao Su, Sihong He, Shuo Han, Haizhao Yang, and Fei Miao. What is the solution for state adversarial multi-agent reinforcement learning? *arXiv preprint arXiv:2212.02705*, 2022.

Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Fast bellman updates for robust MDPs. In *International Conference on Machine Learning*, pages 1979–1988. PMLR, 2018.

Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Partial policy iteration for l1-robust markov decision processes. *Journal of Machine Learning Research*, 22(275):1–46, 2021.

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

Mehdi Jafarnia-Jahromi, Chen-Yu Wei, Rahul Jain, and Haipeng Luo. A model-free learning algorithm for infinite-horizon average-reward MDPs with near-optimal regret. *arXiv preprint arXiv:2006.04354*, 2020.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR, 2020.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pages 5084–5096, 2021.

William Karush. Minima of functions of several variables with inequalities as side conditions. In *Traces and emergence of nonlinear programming*, pages 217–245. Springer, 2013.

David L Kaufman and Andrew J Schaefer. Robust modified policy iteration. *INFORMS Journal on Computing*, 25(3):396–410, 2013.

Michael J Kearns and Satinder P Singh. Finite-sample convergence rates for Q-learning and indirect algorithms. In *Advances in neural information processing systems*, pages 996–1002, 1999.

Olga Klopp, Karim Lounici, and Alexandre B Tsybakov. Robust matrix completion. *Probability Theory and Related Fields*, 169(1-2):523–564, 2017.

Aounon Kumar, Alexander Levine, Tom Goldstein, and Soheil Feizi. Certifying model accuracy under distribution shifts. *arXiv preprint arXiv:2201.12440*, 2022.

Navdeep Kumar, Esther Derman, Matthieu Geist, Kfir Levy, and Shie Mannor. Policy gradient for s-rectangular robust markov decision processes. *arXiv preprint arXiv:2301.13589*, 2023.

Gen Li, Laixi Shi, Yuxin Chen, Yuantao Gu, and Yuejie Chi. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Gen Li, Yuejie Chi, Yuting Wei, and Yuxin Chen. Minimax-optimal multi-agent RL in Markov games with a generative model. *Neural Information Processing Systems*, 2022a.

Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*, 2022b.

Gen Li, Changxiao Cai, Yuxin Chen, Yuting Wei, and Yuejie Chi. Is Q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*, 2023a.

Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *accepted to Operations Research*, 2023b.

Gen Li, Yuling Yan, Yuxin Chen, and Jianqing Fan. Minimax-optimal reward-agnostic exploration in reinforcement learning. *arXiv preprint arXiv:2304.07278*, 2023c.

Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Operations Research*, 72(1):203–221, 2024.

Yan Li, Tuo Zhao, and Guanghui Lan. First-order policy optimization for robust markov decision process. *arXiv preprint arXiv:2209.10579*, 2022c.

A Rupam Mahmood, Dmytro Korenkevych, Gautham Vasan, William Ma, and James Bergstra. Benchmarking reinforcement learning algorithms on real-world robots. In *Conference on robot learning*, pages 561–591. PMLR, 2018.

Shie Mannor, Duncan Simester, Peng Sun, and John N Tsitsiklis. Bias and variance in value function estimation. In *Proceedings of the twenty-first international conference on Machine learning*, page 72, 2004.

Colin McDiarmid et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1): 148–188, 1989.

Janosch Moos, Kay Hansel, Hany Abdulsamad, Svenja Stark, Debora Clever, and Jan Peters. Robust reinforcement learning: A review of foundations and recent advances. *Machine Learning and Knowledge Extraction*, 4(1):276–315, 2022.

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

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

You Qiaoben, Xinning Zhou, Chengyang Ying, and Jun Zhu. Strategically-timed state-observation attacks on deep reinforcement learning agents. In *ICML 2021 Workshop on Adversarial Machine Learning*, 2021.

Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Neural Information Processing Systems (NeurIPS)*, 2021.

Aurko Roy, Huan Xu, and Sebastian Pokutta. Reinforcement learning under model mismatch. *Advances in neural information processing systems*, 30, 2017.

Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

Reazul Hasan Russel, Bahram Behzadian, and Marek Petrik. Optimizing norm-bounded weighted ambiguity sets for robust mdps. *arXiv preprint arXiv:1912.02696*, 2019.

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

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 19967–20025. PMLR, 2022.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. *arXiv preprint arXiv:2305.16589*, 2023.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving Markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pages 5186–5196, 2018.

Elena Smirnova, Elvis Dohmatob, and Jérémie Mary. Distributionally robust reinforcement learning. *arXiv preprint arXiv:1902.08708*, 2019.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3 (1):9–44, 1988.

Aviv Tamar, Shie Mannor, and Huan Xu. Scaling up robust MDPs using function approximation. In *International conference on machine learning*, pages 181–189. PMLR, 2014.

Kai Liang Tan, Yasaman Esfandiari, Xian Yeow Lee, and Soumik Sarkar. Robustifying reinforcement learning agents via action space adversarial training. In *2020 American control conference (ACC)*, pages 3959–3964. IEEE, 2020.

Chen Tessler, Yonathan Efroni, and Shie Mannor. Action robust reinforcement learning and applications in continuous control. In *International Conference on Machine Learning*, pages 6215–6224. PMLR, 2019.

A. B. Tsybakov. *Introduction to nonparametric estimation*, volume 11. Springer, 2009.

J v. Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Martin J Wainwright. Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning. *arXiv preprint arXiv:1905.06265*, 2019.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. A finite sample complexity bound for distributionally robust q-learning. *arXiv preprint arXiv:2302.13203*, 2023.

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

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

Eric M Wolff, Ufuk Topcu, and Richard M Murray. Robust control of uncertain markov decision processes with temporal logic specifications. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 3372–3379. IEEE, 2012.

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34, 2021.

Huan Xu and Shie Mannor. Distributionally robust Markov decision processes. *Mathematics of Operations Research*, 37(2):288–300, 2012.

Zaiyan Xu, Kishan Panaganti, and Dileep Kalathil. Improved sample complexity bounds for distributionally robust reinforcement learning. *arXiv preprint arXiv:2303.02783*, 2023.

Yuling Yan, Gen Li, Yuxin Chen, and Jianqing Fan. The efficacy of pessimism in asynchronous Q-learning. *arXiv preprint arXiv:2203.07368*, 2022.

Yuling Yan, Gen Li, Yuxin Chen, and Jianqing Fan. The efficacy of pessimism in asynchronous q-learning. *IEEE Transactions on Information Theory*, 2023.

Kunhe Yang, Lin Yang, and Simon Du. Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pages 1576–1584. PMLR, 2021.

Wei H Yang. On generalized holder inequality. 1991.

Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Toward theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6): 3223–3248, 2022a.

Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Toward theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6): 3223–3248, 2022b.

Wenhao Yang, Han Wang, Tadashi Kozuno, Scott M Jordan, and Zhihua Zhang. Avoiding model estimation in robust markov decision processes with a generative model. *arXiv preprint arXiv:2302.01248*, 2023.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal offline reinforcement learning via double variance reduction. *arXiv preprint arXiv:2102.01748*, 2021.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems*, 33:21024–21037, 2020a.

Huan Zhang, Hongge Chen, Duane Boning, and Cho-Jui Hsieh. Robust reinforcement learning on state observations with learned optimal adversary. *arXiv preprint arXiv:2101.08452*, 2021.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33, 2020b.

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

6 Other related works

Here we provide additional discussion of related work that could not be fit into the main paper due to space considerations. We limit our discussions to the tabular setting with finite state and action spaces provable RL algorithms.

Classical reinforcement learning with finite-sample guarantees. A recent surge in attention for RL has leveraged the methodologies derived from high-dimensional probability and statistics to analyze RL algorithms in non-asymptotic scenarios. Substantial efforts have been devoted to conducting non-asymptotic sample analyses of standard RL in many settings. Illustrative instances encompass investigations employing Probably Approximately Correct (PAC) bonds in the context of *generative model* settings [Kearns and Singh, 1999, Beck and Srikant, 2012, Li et al., 2022a, Chen et al., 2020, Azar et al., 2013b, Sidford et al., 2018, Agarwal et al., 2020, Li et al., 2023a,b, Wainwright, 2019] and the *online setting* via both in PAC-base or regret-based analyses [Jin et al., 2018, Bai et al., 2019, Li et al., 2021, Zhang et al., 2020b, Dong et al., 2019, Jin et al., 2020, Li et al., 2023c, Jafarnia-Jahromi et al., 2020, Yang et al., 2021] and finally *offline setting* [Rashidinejad et al., 2021, Xie et al., 2021, Yin et al., 2021, Shi et al., 2022, Li et al., 2022b, Jin et al., 2021, Yan et al., 2022].

Robustness in reinforcement learning. Reinforcement learning has had notable achievements but has also exhibited significant limitations, particularly when the learned policy is susceptible to deviations in the deployed environment due to perturbations, model discrepancies, or structural modifications. To address these challenges, the idea of robustness in RL algorithms has been studied. Robustness could concern uncertainty or perturbations across different Markov Decision Processes (MDPs) components, encompassing reward, state, action, and the transition kernel. Moos et al. [2022] gives a recent overview of the different work in this field.

The distributionally robust MDP (RMDP) framework has been proposed [Iyengar, 2005] to enhance the robustness of RL has been proposed. In addition to this work, various other research efforts, including, but not limited to, Zhang et al. [2020a, 2021], Han et al. [2022], Clavier et al. [2022], Qiaoben et al. [2021], explore robustness regarding state uncertainty. In these scenarios, the agent's policy is determined on the basis of perturbed observations generated from the state, introducing restricted noise, or undergoing adversarial attacks. Finally, robustness considerations extend to uncertainty in the action domain. Works such as Tessler et al. [2019], Tan et al. [2020] consider the robustness of actions, acknowledging potential distortions introduced by an adversarial agent.

Given the focus of our work, we provide a more detailed background on progress related to distributionally robust RL. The idea of distributionally robust optimization has been explored within the context of supervised learning [Rahimian and Mehrotra, 2019, Gao, 2020, Duchi and Namkoong, 2018, Blanchet and Murthy, 2019] and has also been extended to distributionally robust dynamic programming and Distributionally Robust Markov Decision Processes (DRMDPs) such as in [Iyengar, 2005, Xu and Mannor, 2012, Wolff et al., 2012, Kaufman and Schaefer, 2013, Ho et al., 2018, Smirnova et al., 2019, Ho et al., 2021, Goyal and Grand-Clement, 2022, Derman and Mannor, 2020, Tamar et al., 2014, Badrinath and Kalathil, 2021]. Despite the considerable attention received, both empirically and theoretically, most previous theoretical analyses in the context of RMDPs adopt an asymptotic perspective [Roy et al., 2017] or focus on planning with exact knowledge of the uncertainty set [Iyengar, 2005, Xu and Mannor, 2012, Tamar et al., 2014]. Many works have focused on the finite-sample performance of verifiable robust Reinforcement Learning (RL) algorithms. These investigations encompass various data generation mechanisms and uncertainty set formulations over the transition kernel. Closely related to our work, various forms of uncertainty sets have been explored, showcasing the versatility of approaches. Divergence such as Kullback-Leibler (KL) divergence is another prevalent choice, extensively studied by Yang et al. [2022a], Panaganti and Kalathil [2022], Zhou et al. [2021], Shi and Chi [2022], Xu et al. [2023], Wang et al. [2023], Blanchet et al. [2023], who investigated the sample complexity of both model-based and model-free algorithms in simulator or offline settings. Xu et al. [2023] considered various uncertainty sets, including those associated with the Wasserstein distance. The introduction of an R-contamination uncertainty set Wang and Zou [2021], has been proposed to tackle a robust Q-learning algorithm for the online setting, with guarantees analogous to standard RL. Finally, the finite-horizon scenario has been studied by Xu et al. [2023], Dong et al. [2022] with finite-sample complexity bounds for (RMDPs) using TV and χ^2 divergence. More broadly, other related topics have been explored, such as the iteration complexity of policy-based methods [Li et al., 2022c, Kumar et al., 2023], and regularization-based robust RL [Yang et al., 2023]. Finally, Badrinath and

Kalathil [2021] examined a general sa-rectangular form of the uncertainty set, proposing a model-free algorithm for the online setting with linear function approximation to address large state spaces.

7 Discussion on hypothesis of Theorems 1 and 3.

- *What norms are included in the Definition 1?* In our upper bound result Theorems 3 and 1, we upper bound the sample complexity for C^2 norms and TV. The set of C^2 smooth norm is very large as it includes all, L_p norm, weighted, rescaled L_p norms for $p \geq 2$. Weighted norms can be useful in practice, to get more weights on dangerous specific states in Robust MDPs formulation such as in Russel et al. [2019]. Moreover, note that our result can generalize to metric or pseudo metric (which are not homogeneous ie $\|\lambda\| = |\lambda| \, \|x\| \, \forall x \in \mathbb{R}^n, \lambda \in \mathbb{R}$) with norms of the form $x \mapsto \phi^{-1}(\sum_{k=1}^{n}, \phi(|x_k|))$ with ϕ a convex incising function such as the norm is still positive, definite positive. Choosing $\phi(x) = x^p$ leads to the L_p norms.
- *Assumptions on γ in Theorems 1 and 3, and Assumptions on γ for lower bound.* When γ is small (e.g., $\gamma \in (0, \frac{1}{2}]$ leads to the effective horizon length is at most 2), the sequential structure almost disappears and is much less of interest for RL community. So people Li et al. [2023b] Yan et al. [2023] usually focus on reasonable range $\gamma \in (c, 1)$ for some small positive constant c, such as $\gamma \in [\frac{1}{2}, 1)$. However, the theorems can be directly extended to a broader range of $\gamma \in (c, 1)$ along with c as small as desired so that almost cover the full range $(0, 1)$.
- *Why final results on s depend on $\hat{\pi}$*

  Theorem 3 is $\hat{\pi}$ data dependent which is randomness-dependent measure. However, taking the minimum of this quantity leads to the same bound as is sa-rectangular, so to illustrate that it is possible to get tighter bounds for s-rectangular with instance-dependent RMDPs, we decide to write also randomness-dependent quantity, while the less tight upper bound is written also in the theorem, taking the first term in the "min" in (21).
- *Why our results are still true for TV?* Theorems 1 and 3 are stated for C^2 smooth norms, however, our result is still true for TV which is not C^2 as in this specific case, the dual of the optimization problem becomes a $1-$dimensional problem. In this case in the main concentration lemma 8, the additional term involving smoothness term denoted C_S is not present and the bound is simpler as is not required this additional term.
- *Why burn-in or sufficiently small ϵ condition is not too restrictive?* The burn-in term in Th. 1 and 3 is proportional to $1/\epsilon$ where the "sample complexity" term is proportional to $1/\epsilon^2$. The smooth term depending on C_S or burn-in is then not too large for sufficiently small ϵcompared to the other term, which will give final the sample complexity.
- *Why this is not extendable to f-divergence currently?* The f-divergence as a distinct family of divergence is beyond the scope of this paper. Current proof for arbitrary norms cannot be directly extended since the key phenomenon of shrinking range of the robust value function has not been verified for f-divergence yet, while it is promising as an interesting future direction.

8 Preliminaries

These quantities appear in the dual formulation of the robust optimization problem and more preciously the dual span semi norm $\mathrm{sp}(.)_*$ note that for L_2, we retrieve the classical mean with the definition of ω) With slight abuse of notation, we denote 0 (resp. 1) as the all-zero (resp. all-one) vector. We then introduce the notation $[T] \coloneqq \{1, \cdots, T\}$ for any positive integer $T > 0$. Then, for two vectors $x = [x_i]_{1 \leq i \leq n}$ and $y = [y_i]_{1 \leq i \leq n}$, the notation $x \leq y$ (resp. $x \geq y$) means $x_i \leq y_i$ (resp. $x_i \geq y_i$) for all $1 \leq i \leq n$. Finally, for any vector x, we overload the notation by letting $x^{\circ 2} = \big[x(s,a)^2\big]_{(s,a) \in \mathcal{S} \times \mathcal{A}}$ (resp. $x^{\circ 2} = \big[x(s)^2\big]_{s \in \mathcal{S}}$), Finally, we drop the subscript $\|.\|$ to write $\mathcal{U}^{\sigma}_{\|.\|}(\cdot) = \mathcal{U}^{\sigma}(\cdot)$ for both sa- and s- rectangular assumptions.

Matrix and Vector Notations. Throughout the analysis, we need to introduce or recall some matrix and vector notations in the following.

- $r \in \mathbb{R}^{SA}$: the reward function vector r (so that $r_{(s,a)} = r(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$).

- $P^0 \in \mathbb{R}^{SA\times S}$: the nominal transition kernel matrix with $P^0_{s,a}$ as the (s,a)-th row.
- $\widehat{P}^0 \in \mathbb{R}^{SA\times S}$: the estimated nomimal transition kernel matrix with $\widehat{P}^0_{s,a}$ as the (s,a)-th row.
- $\Pi^\pi \in \{0,1\}^{S\times SA}$: a projection matrix associated with a given policy π taking the following form:

$$\Pi^\pi = \begin{pmatrix} 1^\top_{\pi(1)} & 0^\top & \cdots & 0^\top \\ 0^\top & 1^\top_{\pi(2)} & \cdots & 0^\top \\ \vdots & \vdots & \ddots & \vdots \\ 0^\top & 0^\top & \cdots & 1^\top_{\pi(S)} \end{pmatrix}, \tag{24}$$

where $1^\top_{\pi(1)}, 1^\top_{\pi(2)}, \dots, 1^\top_{\pi(S)} \in \mathbb{R}^A$ are simplex vector such as

$$1^\top_{\pi(1)} = (\pi(a_1|s_1), \pi(a_A|s_1), ..., \pi(a_A|s_1)).$$

- $P^V \in \mathbb{R}^{SA\times S}$, $\widehat{P}^V \in \mathbb{R}^{SA\times S}$ are the matrices representing the probability transition kernel in the uncertainty set that leads to the worst-case value for any vector $V \in \mathbb{R}^S$. We denote $P^V_{s,a}$ (resp. $\widehat{P}^V_{s,a}$) as the (s,a)-th row of the transition matrix P^V (resp. $\widehat{P}^V$). The (s,a)-th rows of these transition matrices are defined for sa-rectangular assumptions as

$$P^V_{s,a} = \operatorname{argmin}_{\mathcal{P}\in\mathcal{U}^{\mathsf{sa},\sigma}(P^0_{s,a})} \mathcal{P}V, \qquad \text{and} \qquad \widehat{P}^V_{s,a} = \operatorname{argmin}_{\mathcal{P}\in\mathcal{U}^{\mathsf{sa},\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}V. \tag{25a}$$

Moreover, we will use of the following shorthand notation:

$$P^{\pi,V}_{s,a} := P^{V^{\pi,\sigma}}_{s,a} = \operatorname{argmin}_{\mathcal{P}\in\mathcal{U}^{\mathsf{sa},\sigma}(P^0_{s,a})} \mathcal{P}V^{\pi,\sigma}, P^{\pi,\widehat{V}}_{s,a} := P^{\widehat{V}^{\pi,\sigma}}_{s,a} = \operatorname{argmin}_{\mathcal{P}\in\mathcal{U}^{\mathsf{sa},\sigma}(P^0_{s,a})} \mathcal{P}\widehat{V}^{\pi,\sigma}, \tag{25b}$$

$$\widehat{P}^{\pi,V}_{s,a} := \widehat{P}^{V^{\pi,\sigma}}_{s,a} = \operatorname{argmin}_{P\in\mathcal{U}^{\mathsf{sa},\sigma}(\widehat{P}^0_{s,a})} PV^{\pi,\sigma}, \widehat{P}^{\pi,\widehat{V}}_{s,a} := \widehat{P}^{\widehat{V}^{\pi,\sigma}}_{s,a} = \operatorname{argmin}_{P\in\mathcal{U}^{\mathsf{sa},\sigma}(\widehat{P}^0_{s,a})} P\widehat{V}^{\pi,\sigma}. \tag{25c}$$

The corresponding probability transition matrices are denoted by $P^{\pi,V} \in \mathbb{R}^{SA\times S}$, $P^{\pi,\widehat{V}} \in \mathbb{R}^{SA\times S}$, $\widehat{P}^{\pi,V} \in \mathbb{R}^{SA\times S}$ and $\widehat{P}^{\pi,\widehat{V}} \in \mathbb{R}^{SA\times S}$, respectively.

- $P^\pi \in \mathbb{R}^{S\times S}$, $\widehat{P}^\pi \in \mathbb{R}^{S\times S}$, $\underline{P}^{\pi,V} \in \mathbb{R}^{S\times S}$, $\underline{P}^{\pi,\widehat{V}} \in \mathbb{R}^{S\times S}$, $\underline{\widehat{P}}^{\pi,V} \in \mathbb{R}^{S\times S}$ and $\underline{\widehat{P}}^{\pi,\widehat{V}} \in \mathbb{R}^{S\times S}$: six *square* probability transition matrices w.r.t. policy π over the states, namely

$$P^\pi := \Pi^\pi P^0, \qquad \widehat{P}^\pi := \Pi^\pi \widehat{P}^0, \qquad \underline{P}^{\pi,V} := \Pi^\pi P^{\pi,V}, \qquad \underline{P}^{\pi,\widehat{V}} := \Pi^\pi P^{\pi,\widehat{V}},$$
$$\underline{\widehat{P}}^{\pi,V} := \Pi^\pi \widehat{P}^{\pi,V}, \qquad \text{and} \qquad \underline{\widehat{P}}^{\pi,\widehat{V}} := \Pi^\pi \widehat{P}^{\pi,\widehat{V}}. \tag{26}$$

For s-rectangular, we will use the same notation for these transition matrices, removing a subscript for s-rectangular assumptions. Finally, we denote P^π_s as the s-th row of the transition matrix P^π.

- $r_\pi \in \mathbb{R}^S$: a reward restricted to the actions chosen by the policy vector π, $r_\pi = \Pi^\pi r$.
- $\mathrm{Var}_P(V) \in \mathbb{R}^{SA}$: for a given transition kernel $P \in \mathbb{R}^{SA\times S}$ and vector $V \in \mathbb{R}^S$, we denote the (s,a)-th row of $\mathrm{Var}_P(V)$ as

$$\mathsf{Var}_P(s,a) := \mathrm{Var}_{P_{s,a}}(V). \tag{27}$$

8.1 Additional definitions and basic facts

For any norm smooth $\|.\|$ introduced in 1, we define the span semi norm as

Definition 2 (Span semi norm)**.** *Given any norm* $\|\cdot\|$, *we define the span semi norm as:* $\mathrm{sp}(x) = \min_{\omega\in\mathbb{R}} \|v - \omega\mathbf{1}\|$ *and the generalized mean as* $\omega(x) := \arg\min_{\omega\in\mathbb{R}} \|x - \omega\mathbf{1}\|$.

Let vector $P \in \mathbb{R}^{1\times S}$ and vector $V \in \mathbb{R}^S$, we define the variance

$$\mathrm{Var}_P(V) := P(V \circ V) - (PV) \circ (PV). \tag{28}$$

The following lemma bounds the Lipschitz constant of the variance function.

Lemma 1. *(Shi et al. [2023] , Lemma 2) Assuming $0 \leq V_1, V_2 \leq \frac{1}{1-\gamma}$ which obey $\|V_1 - V_2\|_\infty \leq x$, then for $P \in \Delta(S)$, one has*

$$|\mathrm{Var}_P(V_1) - \mathrm{Var}_P(V_2)| \leq \frac{2x}{(1-\gamma)}. \tag{29}$$

Lemma 2. *[Panaganti and Kalathil, 2022, Lemma 6] Consider any $\delta \in (0,1)$. For any fixed policy π and fixed value vector $V \in \mathbb{R}^S$, one has with probability at least $1-\delta$,*

$$\left|\sqrt{\mathrm{Var}_{\widehat{P}^\pi}(V)} - \sqrt{\mathrm{Var}_{P^\pi}(V)}\right| \leq \sqrt{\frac{2\|V\|_\infty^2 \log(\frac{2SA}{\delta})}{N}}1.$$

8.2 Empirical robust MDP $\widehat{\mathcal{M}}_{\mathsf{rob}}$ Bellman equations

We define the robust MDP $\widehat{\mathcal{M}}_{\mathsf{rob}} = \{\mathcal{S}, \mathcal{A}, \gamma, \mathcal{U}^\sigma(\widehat{P}^0), r\}$ based on the estimated nominal distribution $\widehat{P}^0$ in (11). Then, we denote the associated robust value function (resp. robust Q-function) are $\widehat{V}^{\pi,\sigma}$ (resp. $\widehat{Q}^{\pi,\sigma}$). We can notice that that $\widehat{Q}^{\star,\sigma}$ is the unique-fixed point of $\widehat{\mathcal{T}}^\sigma(\cdot)$ (see Lemma 8.3), the empirical robust Bellman operator constructed using $\widehat{P}^0$. Finally, similarly to (9), for $\widehat{\mathcal{M}}_{\mathsf{rob}}$, the Bellman's optimality principle gives the following *robust Bellman consistency equation* (resp. *robust Bellman optimality equation*) for sa-rectangular assumptions:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \widehat{Q}^{\pi,\sigma}(s,a) = r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{sa},\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\pi,\sigma}, \tag{30a}$$

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \widehat{Q}^{\star,\sigma}(s,a) = r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{sa},\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}. \tag{30b}$$

Using matrix notation, we can write the robust Bellman consistency equations as

$$Q^{\pi,\sigma} = r + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{sa},\sigma}(P^0)} \mathcal{P}V^{\pi,\sigma} \quad \text{and} \quad \widehat{Q}^{\pi,\sigma} = r + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{sa},\sigma}(\widehat{P}^0)} \mathcal{P}\widehat{V}^{\pi,\sigma}, \tag{31}$$

which imply

$$\begin{aligned}
V^{\pi,\sigma} &= r_\pi + \gamma \Pi^\pi \inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{sa},\sigma}(P^0)} \mathcal{P}V^{\pi,\sigma} \overset{\text{(i)}}{=} r_\pi + \gamma \underline{P}^{\pi,V} V^{\pi,\sigma}, \\
\widehat{V}^{\pi,\sigma} &= r_\pi + \gamma \Pi^\pi \inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{sa},\sigma}(\widehat{P}^0)} \mathcal{P}\widehat{V}^{\pi,\sigma} \overset{\text{(ii)}}{=} r_\pi + \gamma \underline{\widehat{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma},
\end{aligned} \tag{32}$$

where (i) and (ii) hold by the definitions in (24), (25) and (26). For s-rectangular, we can define the same notation, removing a subscript:

$$\begin{aligned}
V^{\pi,\sigma} &= r_\pi + \gamma \Pi^\pi \inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{s},\tilde{\sigma}}(P^0)} \mathcal{P}V^{\pi,\sigma} \overset{\text{(i)}}{=} r_\pi + \gamma \underline{P}^{\pi,V} V^{\pi,\sigma}, \\
\widehat{V}^{\pi,\sigma} &= r_\pi + \gamma \Pi^\pi \inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{s},\tilde{\sigma}}(\widehat{P}^0)} \mathcal{P}\widehat{V}^{\pi,\sigma} \overset{\text{(ii)}}{=} r_\pi + \gamma \underline{\widehat{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma},.
\end{aligned} \tag{33}$$

8.3 Properties of the robust Bellman operator and dual representation

The robust Bellman operator (cf. (10)) shares the γ-contraction property of the standard Bellman operator as:

[Iyengar, 2005, Theorem 3.2] Given $\gamma \in [0,1)$, the robust Bellman operator $\mathcal{T}^\sigma(\cdot)$ (cf. (10)) is a γ-contraction w.r.t. $\|\cdot\|_\infty$. More formally, for any $Q_1, Q_2 \in \mathbb{R}^{SA}$ s.t. $Q_1(s,a), Q_2(s,a) \in \left[0, \frac{1}{1-\gamma}\right]$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, one has

$$\left\|\mathcal{T}^\sigma(Q_1) - \mathcal{T}^\sigma(Q_2)\right\|_\infty \leq \gamma \left\|Q_1 - Q_2\right\|_\infty. \tag{34}$$

It can be also shown that, $Q^{\star,\sigma}$ is the unique fixed point of $\mathcal{T}^\sigma(\cdot)$ obeying $0 \le Q^{\star,\sigma}(s,a) \le \frac{1}{1-\gamma}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

One of the main contributions is to derive the dual form of optimization problem using arbitrary norms. These lemma take ideas from Iyengar [2005] and are adapted to arbitrary norms and not only TV distance.

Dual equivalence of the robust Bellman operator. Fortunately, the robust Bellman operator can be evaluated efficiently by resorting to its dual formulation, and this idea is central in all proofs for RMPDs. Dual formulation of RMDPs have been introduced in [Iyengar, 2005] but the proof was done uniquely for the TV and the χ^2 case. Before continuing, for any $V \in \mathbb{R}^S$, we denote $[V]_\alpha$ as its clipped version by some non-negative vector α, namely,

$$[V]_\alpha(s) \coloneqq \begin{cases} \alpha, & \text{if } V(s) > \alpha(s), \\ V(s), & \text{otherwise.} \end{cases} \tag{35}$$

Defining the gradient of $P \mapsto \|P\|$ as $\nabla \|P\|$, $\lambda > 0$, a positive scalar and ω is the generalized mean defined as the argmin in the definition of the span semi norm in Def.2, we derive two optimization lemmas.

Lemma 3 (Strong duality using norm $\|\|$ in the sa-rectangular case.)**.** *Consider any probability vector $P \in \Delta(\mathcal{S})$ and any fixed uncertainty level σ, we abbreviate the notation of the uncertainty set $\mathcal{U}^{\mathsf{sa},\sigma}_{\|\cdot\|}(P)$ (cf.* (3)*) as $\mathcal{U}^{\mathsf{sa},\sigma}(P)$. For any vector $V \in \mathbb{R}^S$ obeying $V \ge 0$, recalling the definition of $[V]_\alpha$ in* (35)*, one has*

$$\inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{sa},\sigma}(P)} \mathcal{P}V = \max_{\mu_P^{\lambda,\omega} \in \mathcal{M}_P^{\lambda,\omega}} \left\{ P(V - \mu_P^{\lambda,\omega}) - \sigma \left(\mathrm{sp}((V - \mu_P^{\lambda,\omega}))_* \right) \right\}. \tag{36}$$

$$= \max_{\alpha_P^{\lambda,\omega} \in \mathrm{A}_P^{\lambda,\omega}} \left\{ P\,[V]_{\alpha_P^{\lambda,\omega}} - \sigma \left(\mathrm{sp}([V]_{\alpha_P^{\lambda,\omega}})_* \right) \right\} \tag{37}$$

where $\mathrm{sp}()_*$ *is defined in Def..2. Here, the two auxiliary variational family $\mathrm{A}_P^{\lambda,\omega}, \mathcal{M}_P^{\lambda,\omega}$ are defined as below:*

$$\mathrm{A}_P^{\lambda,\omega} = \{\alpha_P^{\lambda,\omega} : \alpha_P^{\lambda,\omega}(s) = \omega + \lambda |\nabla \|P\|(s) : \lambda > 0, w > 0, P \in \Delta(S), \alpha_P^{\lambda,\omega} \in \left[0, \frac{1}{1-\gamma}\right]^S \} \tag{38}$$

$$\mathcal{M}_P^{\lambda,\omega} = \{\mu_P^{\lambda,\omega} = V - \alpha_P^{\lambda,\omega}, \lambda, \omega \in \mathbb{R}^+, P \in \Delta(S), \mu \in \mathbb{R}^S_+, \mu_P^{\lambda,\omega} = \left[0, \frac{1}{1-\gamma}\right]^S \} \tag{39}$$

$$\tag{40}$$

For L_1 or TV, case , the vector $\alpha_P^{\lambda,\omega}$ reduces to a 1 dimensional scalar such as $\alpha \in [0, 1/(1-\gamma)]$.

Proof.

$$\begin{aligned} \inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{sa},\sigma}(P)} \mathcal{P}V &= \inf_{\{\mathcal{P}: \mathcal{P} \in \Delta_s, \|\mathcal{P} - P\| \le \sigma\}} \sum_{s'} \mathcal{P}(s') V(s') \\ &= PV + \inf_{\{y: \|y\| \le \sigma, 1y = 0, y \ge -P\}} \sum_{s'} y(s') V(s') \end{aligned}$$

where we use the change of variable $y(s') = \mathcal{P}(s') - P(s')$ for all $s' \in \mathcal{S}$. Then the Lagrangian function of the above optimization problem can be written as follows:

$$\begin{aligned} \inf_{\mathcal{P} \in \mathcal{U}^\sigma_{s,a}(P)} \mathcal{P}V =& PV + \sup_{\mu \ge 0, \nu \in \mathbb{R}} \inf_{\{y: \|y\| \le \sigma\}} - \sum_{s'} \mu(s) P(s') + \sum_{s'} (y(s')(V(s') - \mu(s') - \nu) && (41) \\ \overset{(a)}{=}& PV + \sup_{\mu \ge 0, \nu \in \mathbb{R}} - \sum_{s'} \mu(s') P(s') - \sigma \|(V(s') - \mu(s') - \nu \mathbf{1})\|_* && (42) \\ \overset{(b)}{=}& \sup_{\mu \ge 0} P(V - \mu) - \sigma \mathrm{sp}(V - \mu)_* && (43) \end{aligned}$$

where $\mu \in \mathbb{R}_+^{\mathcal{S}}$, $\nu \in \mathbb{R}$ are Lagrangian variables, (a) is true using the equality case of Cauchy-Swartz inequality for dual norm Yang [1991], and (b) is due to is the definition of the span semi-norm (see (8)). The value that maximizes the inner maximization problem in (42) in $\omega(V, \mu)$ is the generalized-mean by definition denoted with abbreviate notation ω. If the norm is differentiable, then we have that the equality (a) comes from the generalized Holder's inequality for arbitrary norms Yang [1991], namely, defining $z = (V - \mu - \omega)$, it satisfies

$$z = \|z\|_* \nabla \|y\| \tag{44}$$

The quantity ν is replaced by the generalized mean for equality in (b) while (44) comes from Yang [1991]. Using complementary slackness Karush [2013]stackness let $\mathcal{B} = \{s \in \mathcal{S} : \mu(s) > 0\}$

$$\forall s \in \mathcal{B} : \quad y^*(s) = -P(s), \tag{45}$$

which leads to the following equality by plugging the previous (45) in (44) and defining $z^* = V - \mu^* - \omega$:

$$\forall s \in \mathcal{B}, \quad z^*(s) = \|z^*\|_* \nabla \|P\| (s) \tag{46}$$

or

$$\forall s \in \mathcal{B}, \quad V(s) - \mu^*(s) = \omega + \lambda \nabla \|P\| (s) \hat{=} \alpha_P^{\lambda,\omega} \tag{47}$$

by letting $\lambda = \|z^*\|_* \in \mathbb{R}^+$. Note that here the hypothesis of 1 are use and especially separability is needed to ensure that for $s \in \mathcal{B}$, $\nabla \|y\| = \nabla \|P\|$ only depend on $P(s)$ and not on other coordinates, which is true form generalized L_p norms. We can remark that $v - \mu^*$ is P dependent, but if P is known, the best μ^* is only determined by one 2 dimensional parameters $\lambda = \|v - \mu^* - \nu\|_*$ and $\omega \in \mathbb{R}^+$. Moreover, when P is fixed, the scalar ω is a constant is fully determined by P, v and μ^*. This is why the quantity defined α_P^λ varies through 2 parameter λ and ω. Given this observation, we can rewrite the optimization problem as :

$$\sup_{\mu \geq 0} P(V - \mu) - \sigma \mathrm{sp}(V - \mu)_* = \sup_{\mu_P^{\lambda,\omega} \in \mathcal{M}_P^{\lambda,\omega}} P(V - \mu_P^{\lambda,\omega}) - \sigma \mathrm{sp}((V - \mu_P^{\lambda,\omega}))_* \tag{48}$$

$$= \sup_{\alpha_P^{\lambda,\omega} \in \mathrm{A}_P^{\lambda,\omega}} P[V]_{\alpha_P^{\lambda,\omega}} - \sigma \mathrm{sp}([V]_{\alpha_P^{\lambda,\omega}})_* \tag{49}$$

where we defined the maximization problem on μ not in $\mathbb{R}^S$ but at the optimal in the variational family denote $\mathcal{M}_P^{\lambda,\omega} = \{v - \alpha_P^{\lambda,\omega}, (\lambda, \omega) \in \mathbb{R}_+^2, P \in \Delta(S)\}$. We can rewrite the optimization problem in terms of α_P with

$$[V]_{\alpha_P^{\lambda,\omega}}(s) := \begin{cases} \alpha_P^{\lambda,\omega}, & \\ V(s), & \text{otherwise.} \end{cases}$$

Contrary to the TV case, α is not a scalar but $\alpha_P^{\lambda,\omega}$ belongs to a variational family only determined by two parameter. Note that this lemma is still true writing subgradient and not gradient of P. As we assume C^2-regularity on norms, the subgradient space of the norm reduce to the singleton of the gradient in our case. C^2 smoothness will be needed in concentration part while it is possible to be more general in optimization lemmas. Note that for TV or L_1, this lemma holds, but the vector $\alpha_P^{\lambda,\omega}$ reduces to a positive scalar denoted α which is equal to $\|v - \mu^*\|_\infty$ according to Iyengar [2005]

□

Lemma 4 (Strong duality for the distance induced by the norm $\|\|$ in the s-rectangular case.)**.** *Consider any probability vector $P^\pi := \Pi^\pi P \in \Delta_s$ for $P \in \Delta(S)^{\mathcal{A}}$, any fixed uncertainty level $\tilde{\sigma}$ and the uncertainty set $\mathcal{U}_{\|.\|}^{\mathsf{s},\tilde{\sigma}}(P)$, we abbreviate the subscript to use $\mathcal{U}^{\mathsf{s},\tilde{\sigma}}(P) := \mathcal{U}_{\|.\|}^{\mathsf{s},\tilde{\sigma}}(P)$. Then for any vector $V \in \mathbb{R}^S$ obeying $V \geq 0$, recalling the definition of $[V]_\alpha$ in* (35)*, one has*

$$\inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{s},\tilde{\sigma}}(P)} \mathcal{P}^\pi V = \sum_a \pi(a|s) \Big(\Big(\max_{\alpha_{P_{sa}}^{\lambda,\omega} \in \mathrm{A}_{P_{sa}}^{\lambda,\omega}} P_{sa}[V]_{\alpha_{P_{sa}}^{\lambda,\omega}} - \tilde{\sigma} \|\pi_s\|_* \mathrm{sp}([V]_{\alpha_{P_{sa}}^{\lambda,\omega}})_* \Big). \tag{50}$$

with the definition of $\mathrm{sp}()_*$ *in 8 and where the variational family* $\mathrm{A}_P^{\lambda,\omega}$ *is defined as :*

$$\mathrm{A}_P^{\lambda,\omega} = \{\alpha \in \Big[0, 1/(1-\gamma)\Big]^S, \alpha = \omega + \lambda|\nabla \|P\|\,| := \alpha_P^{\lambda,\omega}\} \tag{51}$$

$$\tag{52}$$

with ω *is the generalized mean defined as the argmin in the definition of the span semi norm in 2 and* λ, ω *a positive scalar. Moreover, for* L_1 *or* TV*, case, the vector* $\alpha_P^{\lambda,\omega}$ *reduces to a* 1 *dimensional scalar such as* $\alpha \in [0, 1/(1-\gamma)]$.

In the proof of the previous lemma, we decompose this problem s-rectangular radius $\tilde{\sigma}$ into sa-rectangular sub-problem with respectively radius σ_{sa}.

Proof.

$$\begin{aligned}
&\inf_{\mathcal{P}^\pi \in \mathcal{U}^{\mathsf{s},\tilde{\sigma}}(P^\pi)} \mathcal{P}^\pi V = \inf_{\{\sigma_{sa}: \|\sigma_{sa}\| \leq \tilde{\sigma}\}} \inf_{\mathcal{P}' \in \mathcal{U}^{\mathsf{sa},\sigma}(P_{sa})} \sum_a \pi(a|s) \mathcal{P}' V \\
&\overset{(a)}{=} \sum_a \pi(a|s) P_{sa} V + \min_{\{\sigma_{sa}: \|\sigma_{sa}\| \leq \tilde{\sigma}\}} \sum_a \pi(a|s) \min_{\{y: \|y\| \leq \sigma_{sa}, 1y=0, y \geq -P_{sa}\}} \sum_{s'} y(s') V
\end{aligned}$$

where we use the change of variable $y(s') = \mathcal{P}_{sa}(s') - P_{sa}(s')$ in (a). Then we case use the previous lemma for sa rectangular assumption, Lemma 3. Then,

$$\begin{aligned}
&\min_{\{\sigma_{sa}: \|\sigma_{sa}\| \leq \tilde{\sigma}\}} \sum_a \pi(a|s) \min_{\{y, \|y\| \leq \sigma_{s,a}, 1y=0, y \geq -P_{sa}\}} \sum_{s'} y(s') V \\
&= \min_{\{\sigma_{sa}: \|\sigma_{sa}\| \leq \tilde{\sigma}\}} \sum_a \pi(a|s) \max_{\mu \geq 0} \Big(- P_{sa}\mu - \sigma_{sa} \mathrm{sp}(V-\mu)_* \Big) \\
&= \left(\sum_a \pi(a|s) \max_{\mu \geq 0} \Big\{ (-P_{sa}\mu) - \max_{\{\sigma_{sa}: \|\sigma_{sa}\| \leq \tilde{\sigma}\}} \sum_a \pi(a|s) \sigma \mathrm{sp}(V-\mu)_* \Big\} \right) \\
&= \sum_a \pi(a|s) \max_{\mu \geq 0} \Big\{ (-P_{sa}\mu) - \tilde{\sigma} \|\pi_s\|_* \mathrm{sp}(V-\mu)_* \Big\}.
\end{aligned}$$

We can exchange the min and the max as we get concave-convex problems in σ and μ in the second line according to minimax theorem [v. Neumann, 1928] and using Cauchy Swartz inequality which is attained in the last equality. Finally, we obtain:

$$\begin{aligned}
\inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{s},\tilde{\sigma}}(P)} \mathcal{P}^\pi V &= \sum_a \pi(a|s) \Big(\max_{\mu \geq 0} P_{sa}(V-\mu) - \tilde{\sigma} \|\pi_s\|_* \mathrm{sp}(V-\mu)_* \Big) \\
&\overset{(a)}{=} \sum_a \pi(a|s) \Big(\max_{\alpha_{P_{sa}}^{\lambda,\omega} \in \mathrm{A}_{P_{sa}}^{\lambda,\omega}} P_{sa}[V]_{\alpha_{P_{sa}}^{\lambda,\omega}} - \tilde{\sigma} \|\pi_s\|_* \mathrm{sp}([V]_{\alpha_{P_{sa}}^{\lambda,\omega}})_* \Big)
\end{aligned}$$

where in (a) we use the previous lemma for $sa-$ rectangular case. Note that as we are using sa-rectangular case, for TV or L_1, this lemma holds, but the vector α_P^λ reduces to a positive scalar denoted α which is equal to $\|v - \mu^*\|_\infty$. (See also Iyengar [2005]).

□

9 Proof of the upper bound : Theorem 1 and 3

9.1 Technical lemmas

We begin with a key lemma concerning the dynamic range of the robust value function $V^{\pi,\sigma}$ (cf. (7)), which produces tighter control when σ is large; the proof is deferred to Appendix 9.3.1. This lemma allows tighter control compared to Clavier et al. [2023].

Lemma 5. *In $sa-$rectangular case (see* (3)*), for any nominal transition kernel $P \in \mathbb{R}^{SA \times S}$, any fixed uncertainty level σ, and any policy π, its corresponding robust value function $V^{\pi,\sigma}$ (cf.* (7)*) satisfies*

$$\mathrm{sp}(V^{\pi,\sigma})_\infty \leq \frac{1}{\gamma \max\{1-\gamma, C_g \sigma\}} \tag{53}$$

where $C_g = 1/(\min_s \|e_s\|)$ is a geometric constant depending on the geometry of the norm. For example, for L_p, norms $p \geq 1$, $C_g \geq 1$ which reduce the sample complexity. In s-rectangular case, we obtain a slightly different lemma because of the dependency on π.

Lemma 6. *The infinite span semi norm can be controlled as follows for every s in s-rectanuglar case (See* (5)*):*

$$\mathrm{sp}(V^{\pi,\sigma})_\infty \leq \frac{1}{\gamma \max\{1-\gamma, \|\pi_s\|_* C_g \tilde{\sigma}\}} \leq \frac{1}{\gamma \max\{1-\gamma, \min_s \|\pi_s\|_* C_g \tilde{\sigma}\}} \tag{54}$$

where $C_g = \frac{1}{\min_s \|e_s\|}$ is a geometric constant depending on the geometry of the norm. These lemmas are required to get tight bounds for the sample complexity. The main difference between sa- and s-rectangular case is that we have an extra dependency on $\|\pi_s\|_*$, which represents how stochastic the policy can be in s rectangular MDPs.

Lemma 7. *Consider an MDP with transition kernel matrix P and reward function $0 \leq r \leq 1$. For any policy π and its associated state transition matrix $P_\pi := \Pi^\pi P$ and value function $0 \leq V^{\pi,P} \leq \frac{1}{1-\gamma}$ (cf.* (1)*), one has for sa- and s- rectangular assumptions.*

$$(I - \gamma P_\pi)^{-1} \sqrt{\mathrm{Var}_{P_\pi}(V^{\pi,P})} \leq \sqrt{\frac{8}{\gamma^2(1-\gamma)^2} \mathrm{sp}(V^{\pi,P})_\infty} 1.$$

See 9.3.7 for the proof

9.2 Proof of Theorem 1 and Theorem 3

The first decomposition of the proof of Theorem 1 and Theorem 3 Agarwal et al. [2020] while the argument needs essential adjustments in order to adapt to the robustness setting. One has by assumptions using any planner in empirical RMDPs :

$$\left\| \widehat{V}^{\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma} \right\|_\infty \leq \varepsilon_{\mathsf{opt}}, \tag{55}$$

using previous inequality, performance gap $\left\| V^{\star,\sigma} - V^{\widehat{\pi},\sigma} \right\|_\infty$, can be upper bounded using 3 steps.

First step: subdivide the performance gap in 3 terms. We recall the definition of the optimal robust policy $\pi^\star$ with regard to $\mathcal{M}_{\mathsf{rob}}$ and the optimal robust policy $\widehat{\pi}^\star$, the optimal robust value function $\widehat{V}^{\star,\sigma}$ (resp. robust value function $\widehat{Q}^{\pi,\sigma}$) w.r.t. $\widehat{\mathcal{M}}_{\mathsf{rob}}$. Then, the performance gap $V^{\star,\sigma} - V^{\widehat{\pi},\sigma}$ can be decomposed in one optimization term and two statistical error terms

$$\begin{aligned}
V^{\star,\sigma} - V^{\widehat{\pi},\sigma} &= \left(V^{\pi^\star,\sigma} - \widehat{V}^{\pi^\star,\sigma}\right) + \left(\widehat{V}^{\pi^\star,\sigma} - \widehat{V}^{\widehat{\pi}^\star,\sigma}\right) + \left(\widehat{V}^{\widehat{\pi}^\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma}\right) + \left(\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\right) \\
&\overset{\text{(i)}}{\leq} \left(V^{\pi^\star,\sigma} - \widehat{V}^{\pi^\star,\sigma}\right) + \left(\widehat{V}^{\widehat{\pi}^\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma}\right) + \left(\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\right) \\
&\overset{\text{(ii)}}{\leq} \left(V^{\pi^\star,\sigma} - \widehat{V}^{\pi^\star,\sigma}\right) + \varepsilon_{\mathsf{opt}} + \left(\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\right)
\end{aligned} \tag{56}$$

where (i) holds by $\widehat{V}^{\pi^\star,\sigma} - \widehat{V}^{\widehat{\pi}^\star,\sigma} \leq 0$ since $\widehat{\pi}^\star$ is the robust optimal policy for $\widehat{\mathcal{M}}_{\mathsf{rob}}$, and (ii) comes from (55) and definition of optimization error. The proof aims to control the last remaining terms in (56) using concentration theory and sufficiently big number of step N. To do so, we will consider a more general term $\widehat{V}^{\pi,\sigma} - V^{\pi,\sigma}$ for any policy π even if control of these two terms slightly differ at the end. Using (32), it holds that for both sa- and s-rectangular assumptions:

$$\begin{aligned}
\widehat{V}^{\pi,\sigma} - V^{\pi,\sigma} &= r_\pi + \gamma \underline{\widehat{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \left(r_\pi + \gamma \underline{P}^{\pi,V} V^{\pi,\sigma}\right) \\
&= \left(\gamma \underline{\widehat{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \gamma \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}\right) + \left(\gamma \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \gamma \underline{P}^{\pi,V} V^{\pi,\sigma}\right) \\
&\overset{\text{(i)}}{\leq} \gamma \left(\underline{P}^{\pi,V} \widehat{V}^{\pi,\sigma} - \underline{P}^{\pi,V} V^{\pi,\sigma}\right) + \left(\gamma \underline{\widehat{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \gamma \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}\right),
\end{aligned}$$

where (i) holds because $\underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} \leq \underline{P}^{\pi,V} \widehat{V}^{\pi,\sigma}$ because of the optimality of $\underline{P}^{\pi,\widehat{V}}$ (see. (25)). Factorizing terms leads to the following equation

$$\widehat{V}^{\pi,\sigma} - V^{\pi,\sigma} \leq \gamma \left(I - \gamma \underline{P}^{\pi,V}\right)^{-1} \left(\underline{\widehat{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}\right). \tag{57}$$

In the same manner, we can also obtain a lower bound of this quantity:

$$\begin{aligned}
\widehat{V}^{\pi,\sigma} - V^{\pi,\sigma} &= r_\pi + \gamma \underline{\widehat{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \left(r_\pi + \gamma \underline{P}^{\pi,V} V^{\pi,\sigma}\right) \\
&= \left(\gamma \underline{\widehat{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \gamma \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}\right) + \left(\gamma \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \gamma \underline{P}^{\pi,V} V^{\pi,\sigma}\right) \\
&\geq \gamma \left(\underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \underline{P}^{\pi,\widehat{V}} V^{\pi,\sigma}\right) + \left(\gamma \underline{\widehat{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \gamma \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}\right) \\
&\geq \gamma \left(I - \gamma \underline{P}^{\pi,\widehat{V}}\right)^{-1} \left(\underline{\widehat{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}\right).
\end{aligned} \tag{58}$$

Using both (57) and (58), we obtain infinite norm control:

$$\begin{aligned}
\left\|\widehat{V}^{\pi,\sigma} - V^{\pi,\sigma}\right\|_\infty \leq \gamma \max \Bigg\{ &\left\| \left(I - \gamma \underline{P}^{\pi,V}\right)^{-1} \left(\underline{\widehat{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}\right) \right\|_\infty, \\
&\left\| \left(I - \gamma \underline{P}^{\pi,\widehat{V}}\right)^{-1} \left(\underline{\widehat{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}\right) \right\|_\infty \Bigg\}.
\end{aligned} \tag{59}$$

By decomposing the error in a symmetric way, he have

$$\begin{aligned}
\left\|\widehat{V}^{\pi,\sigma} - V^{\pi,\sigma}\right\|_\infty \leq \gamma \max \Bigg\{ &\left\| \left(I - \gamma \underline{\widehat{P}}^{\pi,V}\right)^{-1} \left(\underline{\widehat{P}}^{\pi,V} V^{\pi,\sigma} - \underline{P}^{\pi,V} V^{\pi,\sigma}\right) \right\|_\infty, \\
&\left\| \left(I - \gamma \underline{\widehat{P}}^{\pi,\widehat{V}}\right)^{-1} \left(\underline{\widehat{P}}^{\pi,V} V^{\pi,\sigma} - \underline{P}^{\pi,V} V^{\pi,\sigma}\right) \right\|_\infty \Bigg\}.
\end{aligned} \tag{60}$$

Armed with these inequalities, we can use concentration inequalities to upper bound the two remaining terms $\left\|\widehat{V}^{\pi^\star,\sigma} - V^{\pi^\star,\sigma}\right\|_\infty$ and $\left\|\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\right\|_\infty$ in (56). Taking $\pi = \widehat{\pi}$, applying (59) leads to

$$\begin{aligned}
\left\|\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\right\|_\infty \leq \gamma \max \Bigg\{ &\left\| \left(I - \gamma \underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1} \left(\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}} \widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}} \widehat{V}^{\widehat{\pi},\sigma}\right) \right\|_\infty, \\
&\left\| \left(I - \gamma \underline{P}^{\widehat{\pi},V}\right)^{-1} \left(\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}} \widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}} \widehat{V}^{\widehat{\pi},\sigma}\right) \right\|_\infty \Bigg\}.
\end{aligned} \tag{61}$$

Finally, $\pi = \pi^\star$, applying (60) gives us

$$\begin{aligned}
\left\|\widehat{V}^{\pi^\star,\sigma} - V^{\pi^\star,\sigma}\right\|_\infty \leq \gamma \max \Bigg\{ &\left\| \left(I - \gamma \underline{\widehat{P}}^{\pi^\star,V}\right)^{-1} \left(\underline{\widehat{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\right) \right\|_\infty, \\
&\left\| \left(I - \gamma \underline{\widehat{P}}^{\pi^\star,\widehat{V}}\right)^{-1} \left(\underline{\widehat{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\right) \right\|_\infty \Bigg\}.
\end{aligned} \tag{62}$$

Note that to control $\left\|\widehat{V}^{\pi^\star,\sigma} - V^{\pi^\star,\sigma}\right\|_\infty$, we use decomposition not depending on $\widehat{\pi}$ for value function as $V^{\pi^\star,\sigma}$ is deterministic and fixed, allowing use of classical concentration analysis tools. This decomposition is the same for both sa-rectangular and s-rectangular case.

Second step: bound first term and second term in (62) **to control** $\|\widehat{V}^{\pi^\star,\sigma} - V^{\pi^\star,\sigma}\|_\infty$ To control the two terms in (62), we use lemma 8 based Bernstein's concentration argument and whose proof is in Appendix 9.3.3.

Lemma 8. *For both sa− and s-rectangular setting, consider any $\delta \in (0,1)$, with probability $1-\delta$, it holds:*

$$\left|\widehat{\underline{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\right| \leq 2\sqrt{\frac{L}{N}}\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})} + \frac{3LC_S\|1\|_*}{N(1-\gamma)} \tag{63}$$

with $L = 2\log(18\|1\|_* SAN/\delta)$ and where $\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})$ is defined in (27)*. Moreover, for the specific case of TV, this lemma is true without the smoothness term $\frac{3LC_S\|1\|_*}{N(1-\gamma)}$.*

Armed with the above lemma, now we control the **first term** on the right-hand side of (62) as follows:

$$\begin{aligned}
&\left(I - \gamma\widehat{\underline{P}}^{\pi^\star,V}\right)^{-1}\left(\widehat{\underline{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\right) \\
&\overset{\text{(a)}}{\leq} \left(I - \gamma\widehat{\underline{P}}^{\pi^\star,V}\right)^{-1}\left\|\widehat{\underline{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\right\|_\infty \\
&\overset{\text{(b)}}{\leq} \left(I - \gamma\widehat{\underline{P}}^{\pi^\star,V}\right)^{-1}\left(2\sqrt{\frac{L}{N}}\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})} + \frac{3LC_S\|1\|_*}{N(1-\gamma)}\right) \\
&\leq \left(I - \gamma\widehat{\underline{P}}^{\pi^\star,V}\right)^{-1}\frac{3LC_S\|1\|_*}{N(1-\gamma)}1 + \underbrace{2\sqrt{\frac{L}{N}}\left(I - \gamma\widehat{\underline{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\mathrm{Var}_{\widehat{\underline{P}}^{\pi^\star,V}}(V^{\star,\sigma})}}_{=:\mathcal{R}_1} \\
&\quad + \underbrace{2\sqrt{\frac{L}{N}}\left(I - \gamma\widehat{\underline{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\left|\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma}) - \mathrm{Var}_{\widehat{\underline{P}}^{\pi^\star,V}}(V^{\star,\sigma})\right|}}_{=:\mathcal{R}_2} \\
&\quad + \underbrace{2\sqrt{\frac{L}{N}}\left(I - \gamma\widehat{\underline{P}}^{\pi^\star,V}\right)^{-1}\left(\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})} - \sqrt{\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})}\right)}_{=:\mathcal{R}_3},
\end{aligned} \tag{64}$$

where (a) holds as the matrix $\left(I - \gamma\widehat{\underline{P}}^{\pi^\star,V}\right)^{-1}$ is positive definite, (b) holds due to Lemma 8, and the last point holds from the following decomposition for variance and triangular inequality

$$\begin{aligned}
&\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})} = \left(\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})} - \sqrt{\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})}\right) + \sqrt{\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})} \\
&\leq \left(\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})} - \sqrt{\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})}\right) \\
&+ \sqrt{\left|\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma}) - \mathrm{Var}_{\widehat{\underline{P}}^{\pi^\star,V}}(V^{\star,\sigma})\right|} + \sqrt{\mathrm{Var}_{\widehat{\underline{P}}^{\pi^\star,V}}(V^{\star,\sigma})}.
\end{aligned}$$

Finally, the fact that $\widehat{\underline{P}}^{\pi^\star,V}$ is a stochastic matrix, so

$$\left(I - \gamma\widehat{\underline{P}}^{\pi^\star,V}\right)^{-1}1 = \left(I + \sum_{t=1}^{\infty}\gamma^t\left(\widehat{\underline{P}}^{\pi^\star,V}\right)^t\right)1 \leq \frac{1}{1-\gamma}1. \tag{65}$$

Armed with these inequalities, the three terms $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$ in (64) can be controlled separately.

- Consider $\mathcal{R}_1$. We first introduce the following lemma, whose proof is postponed to Appendix 9.3.4.

Lemma 9. *Consider any $\delta \in (0,1)$. With probability at least $1-\delta$, one has*

$$\left(I - \gamma \widehat{\underline{P}}^{\pi^\star, V}\right)^{-1} \sqrt{\mathrm{Var}_{\widehat{\underline{P}}^{\pi^\star,V}}(V^{\star,\sigma})} \qquad \leq 4\sqrt{\frac{\left(1+\left(\sqrt{\frac{L}{(1-\gamma)^2 N}} + \frac{C_S \|1\|_* L}{N(1-\gamma)}\right)\right)}{\gamma^3 (1-\gamma)^2 \max\{1-\gamma, C_g \sigma\}}} 1$$

$$\leq 4\sqrt{\frac{\left(1+\left(\sqrt{\frac{L}{(1-\gamma)^2 N}} + \frac{C_S \|1\|_* L}{N(1-\gamma)}\right)\right)}{\gamma^3 (1-\gamma)^3}} 1$$

with $L = 2\log(\frac{18\|1\|_* SAN}{\delta})$ in the sa-rectangular case. In the s-rectangular case, it holds:

$$\left(I - \gamma \widehat{\underline{P}}^{\pi^\star, V}\right)^{-1} \sqrt{\mathrm{Var}_{\widehat{\underline{P}}^{\pi^\star,V}}(V^{\star,\sigma})} \leq \qquad \leq 4\sqrt{\frac{\left(1+\left(\sqrt{\frac{L}{(1-\gamma)^2 N}} + \frac{C_S \|1\|_* L}{N(1-\gamma)}\right)\right)}{\gamma^3 (1-\gamma)^2 \max\{1-\gamma, C_g \tilde{\sigma} \min_s \|\pi_s\|_*\}}} 1$$

$$\leq 4\sqrt{\frac{\left(1+\left(\sqrt{\frac{L}{(1-\gamma)^2 N}} + \frac{C_S \|1\|_* L}{N(1-\gamma)}\right)\right)}{\gamma^3 (1-\gamma)^3}} 1$$

Using Lemma 9 and inserting back to (64) gives in sa-rectangular case

$$\begin{aligned}
\mathcal{R}_1 &= 2\sqrt{\frac{L}{N}} \left(I - \gamma \widehat{\underline{P}}^{\pi^\star, V}\right)^{-1} \sqrt{\mathrm{Var}_{\widehat{\underline{P}}^{\pi^\star,V}}(V^{\star,\sigma})} \\
&\leq 8\sqrt{\frac{L}{\gamma^3(1-\gamma)^2 \max\{1-\gamma, C_g\sigma\} N} \left(1 + \sqrt{\frac{L}{(1-\gamma)^2 N}} + \frac{C_S \|1\|_* L}{N(1-\gamma)}\right)} 1. \qquad (66)
\end{aligned}$$

- Consider $\mathcal{R}_2$. First, denote $V' := V^{\star,\sigma} - \eta 1$ $\eta \in \mathbb{R}$, by Lemma 5, we have for any π,

$$0 \leq \min_\eta \|V\|_\infty - \eta 1 \leq \frac{1}{\gamma \max\{1-\gamma, C_g \sigma\}}. \qquad (67)$$

for sa-rectangular case or in s-rectangular we obtain

$$0 \leq \min_\eta \|V - \eta 1\|_\infty \leq \frac{1}{\gamma \max\{1-\gamma, \tilde{\sigma} C_g \|\pi_s\|_*\}} \qquad (68)$$

by the definition of the span semi norm. Moreover, we can use Holder with L_1 and L_∞ we have for both sa and s-rectangular case to as it holds that:

$$\begin{aligned}
&\left|\mathsf{Var}_{\widetilde{P}_{s,a}}(V^{\star,\sigma}) - \mathsf{Var}_{P_{s,a}}(V^{\star,\sigma})\right| = \left|\mathsf{Var}_{\widetilde{P}_{s,a}}(V') - \mathsf{Var}_{P_{s,a}}(V')\right| \\
&\leq \left\|\widetilde{P}_{s,a} - P_{s,a}\right\|_1 \left\|V'\right\|_\infty^2 \overset{a}{\leq} \frac{\sigma_1}{(\gamma^2(\max{(1-\gamma)}, C_g\sigma)^2} \\
&\leq \frac{1}{\gamma^2 \max\{(1-\gamma), \sigma C_g\}} \qquad (69)
\end{aligned}$$

In the first inequality, we use $\left\|V'\right\|_\infty^2 = \left\|V'^2\right\|_\infty$ and and we use Lemma 5 in (a) where $C_g\sigma = \sigma_1$.

With the same arguments for s-rectangular, we obtain for $V' := V^{\star,\sigma} - \eta 1$ $\eta \in \mathbb{R}$,

$$\begin{aligned}
&\left|\Pi^{\pi^\star}\left(\mathsf{Var}_{\widetilde{P}_s}(V^{\star,\sigma}) - \mathsf{Var}_{P_s}(V^{\star,\sigma})\right)\right| = \left|\Pi^{\pi^\star}\left(\mathsf{Var}_{\widetilde{P}_s}(V') - \mathsf{Var}_{P_s}(V')\right)\right| \\
&\leq \sum_a \pi(a|s)(\widetilde{P}_s(s',a) - P_s(s',a))V(s')^2 \qquad (70) \\
&\overset{a}{\leq} \|V'\|_\infty^2 \sum_a \pi(a|s)(\widetilde{P}_s(s',a) - P_s(s',a)) \overset{b}{\leq} \|V'\|_\infty^2 \tilde{\sigma} \|\pi_s\|_* \qquad (71) \\
&\overset{c}{\leq} \frac{\tilde{\sigma} C_g \|\pi_s^*\|_* \left\|V'\right\|_\infty}{\gamma \|\pi_s^*\|_* \tilde{\sigma} C_g} 1 \leq \frac{\left\|V'\right\|}{\gamma} 1. \qquad (72)
\end{aligned}$$

where where (a) and (b) comes Cauchy Swartz inequality, , (c) comes lemma 6. Then, taking the sup over s in the previous equations, it holds

$$
\begin{aligned}
\left|\Pi^{\pi^\star}\left(\mathsf{Var}_{\widetilde{P}_s}(V^{\star,\sigma}) - \mathsf{Var}_{P_s}(V^{\star,\sigma})\right)\right| &\leq \frac{\inf_{\eta\in\mathbb{R}^+}\left\|V-\eta 1'\right\|}{\gamma}1 && (73)\\
&\leq \frac{1}{\gamma^2\tilde{\sigma}\min_s\|\pi_s^*\|_* C_g}1. && (74)
\end{aligned}
$$

Applying the previous inequality, it holds in sa-rectangular case:

$$
\begin{aligned}
\mathcal{R}_2 &= 2\sqrt{\frac{L}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\left|\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma}) - \mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}(V^{\star,\sigma})\right|}\\
&= 2\sqrt{\frac{L}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\left|\Pi^{\pi^\star}\left(\mathrm{Var}_{\widehat{P}^0}(V^{\star,\sigma}) - \mathrm{Var}_{\widehat{P}^{\pi^\star,V}}(V^{\star,\sigma})\right)\right|}\\
&\leq 2\sqrt{\frac{L}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\left\|\mathrm{Var}_{\widehat{P}^0}(V^{\star,\sigma}) - \mathrm{Var}_{\widehat{P}^{\pi^\star,V}}(V^{\star,\sigma})\right\|_\infty 1}\\
&\leq 2\sqrt{\frac{L}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\frac{1}{\gamma^2\max\{1-\gamma, C_g\sigma\}}}1 && (75)\\
&\leq 4\sqrt{\frac{L}{\gamma^2(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}N}}1, && (76)
\end{aligned}
$$

where the last inequality uses $\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}1 \leq \frac{1}{1-\gamma}1$ (cf. (65)). for sa-rectangular
In the s-rectangular case, we obtain a different result as

$$
\begin{aligned}
\mathcal{R}_2 &= 2\sqrt{\frac{L}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\left|\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma}) - \mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}(V^{\star,\sigma})\right|}\\
&= 2\sqrt{\frac{L}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\left|\Pi^{\pi^\star}\left(\mathrm{Var}_{\widehat{P}^0}(V^{\star,\sigma}) - \mathrm{Var}_{\widehat{P}^{\pi^\star,V}}(V^{\star,\sigma})\right)\right|}\\
&\leq 2\sqrt{\frac{L}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\frac{1}{\gamma^2\max\{1-\gamma,\min_s\|\pi_s^*\|_\infty C_g\tilde{\sigma}\}}}1 && (77)\\
&\leq 2\sqrt{\frac{L}{\gamma^2(1-\gamma)^2\max\{1-\gamma,\min_s\|\pi_s^*\|_\infty\tilde{\sigma}C_g\}N}}1, && (78)
\end{aligned}
$$

- Consider $\mathcal{R}_3$. The following lemma plays an important role.
  Applying Lemma 2 and using $\pi=\pi^\star$ and $V=V^{\star,\sigma}$, it holds

$$
\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})} - \sqrt{\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})} \leq \sqrt{\frac{2\|V^{\star,\sigma}\|_\infty^2\log(\frac{2SA}{\delta})}{N}}1,
$$

which can be inserted in (64) to gives

$$
\begin{aligned}
\mathcal{R}_3 &= 2\sqrt{\frac{L}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\left(\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})} - \sqrt{\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})}\right)\\
&\leq \frac{4}{(1-\gamma)}\frac{\log(\frac{SAN}{\delta})\|[V^{\star,\sigma}\|_\infty}{N}1 \leq \frac{4L}{(1-\gamma)^2N}1, && (79)
\end{aligned}
$$

where the last line uses $\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}1 \leq \frac{1}{1-\gamma}1$ (cf. (65)).

Finally, inserting the results of $\mathcal{R}_1$ in (66), $\mathcal{R}_2$ in (78), $\mathcal{R}_3$ in (79), and (65) back into (64) gives

$$\begin{aligned}
&\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\left(\underline{\widehat{P}}^{\pi^\star,V}V^{\pi^\star,\sigma}-\underline{P}^{\pi^\star,V}V^{\pi^\star,\sigma}\right) \qquad (80)\\
&\leq 8\sqrt{\frac{L}{\gamma^3(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}N}\left(1+\sqrt{\frac{L}{(1-\gamma)^2N}}+\frac{C_S\left\|1\right\|_* L}{N(1-\gamma)}\right)}1+\frac{3LC_S\left\|1\right\|_*}{N(1-\gamma)^2}1\\
&\quad+2\sqrt{\frac{2L}{\gamma^2(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}N}}1+\frac{4L}{(1-\gamma)^2N}1\\
&\leq 10\sqrt{\frac{2L}{\gamma^3(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}N}\left(1+\sqrt{\frac{L}{(1-\gamma)^2N}}+\frac{C_S\left\|1\right\|_* L}{N(1-\gamma)}\right)}1+\frac{4L}{(1-\gamma)^2N}1+\frac{3LC_S\left\|1\right\|_*}{N(1-\gamma)^2}1\\
&\leq 160\sqrt{\frac{L(1+\frac{C_S\|1\|_*}{N(1-\gamma)})}{(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}N}}1+\frac{7LC_S\left\|1\right\|_*}{N(1-\gamma)^2}1, \qquad (81)
\end{aligned}$$

where the last inequality holds by the fact $\gamma \geq \frac{1}{4}$ and letting $N \geq \frac{L}{(1-\gamma)^2}$. We have the same result for s-rectangular, replacing, $\max\{1-\gamma, C_g\sigma\}$ by $\max\{1-\gamma, \min_s \|\pi_s^*\|_* \tilde{\sigma} C_g\}$.

Now we are ready to control **second term in** (62) to control $\|\widehat{V}^{\pi^\star,\sigma}-V^{\pi^\star,\sigma}\|_\infty$. To proceed, applying Lemma 8 on the second term of the right-hand side of (62) leads to

$$\begin{aligned}
&\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,\widehat{V}}\right)^{-1}\left(\underline{\widehat{P}}^{\pi^\star,V}V^{\pi^\star,\sigma}-\underline{P}^{\pi^\star,V}V^{\pi^\star,\sigma}\right)\\
&\leq \left(I-\gamma\underline{\widehat{P}}^{\pi^\star,\widehat{V}}\right)^{-1}\left(2\sqrt{\frac{L}{N}}\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})}+\frac{3LC_S\left\|1\right\|_*}{N(1-\gamma)}\right)\\
&\leq \left(I-\gamma\underline{\widehat{P}}^{\pi^\star,\widehat{V}}\right)^{-1}\frac{L'C_S\left\|1\right\|_*}{N(1-\gamma)}+\underbrace{2\sqrt{\frac{L}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,\widehat{V}}}(\widehat{V}^{\pi^\star,\sigma})}}_{=:\mathcal{R}_4}\\
&\quad\underbrace{2\sqrt{\frac{L}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,\widehat{V}}\right)^{-1}\left(\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,\widehat{V}}}(V^{\pi^\star,\sigma}-\widehat{V}^{\pi^\star,\sigma})}\right)}_{=:\mathcal{R}_5}\\
&\quad+\underbrace{2\sqrt{\frac{L}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,\widehat{V}}\right)^{-1}\left(\sqrt{\left|\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})-\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,\widehat{V}}}(V^{\star,\sigma})\right|}\right)}_{=:\mathcal{R}_6}\\
&\quad+\underbrace{2\sqrt{\frac{L}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,\widehat{V}}\right)^{-1}\left(\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})}-\sqrt{\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})}\right)}_{=:\mathcal{R}_7}. \qquad (82)
\end{aligned}$$

We now bound the above four terms $\mathcal{R}_4, \mathcal{R}_5, \mathcal{R}_6, \mathcal{R}_7$ separately.

- Using Lemma 7 with $P = \widehat{P}^{\pi^\star,\widehat{V}}$, $\pi = \pi^\star$ and $V = \widehat{V}^{\pi^\star,\sigma}$ which follow $\widehat{V}^{\pi^\star,\sigma} = r_{\pi^\star} + \gamma\underline{\widehat{P}}^{\pi^\star,\widehat{V}}\widehat{V}^{\pi^\star,\sigma}$, and in view of (65), the term $\mathcal{R}_4$ in (82) can be controlled as follows:

$$\begin{aligned}
\mathcal{R}_4 &= 2\sqrt{\frac{L}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,\widehat{V}}}(\widehat{V}^{\pi^\star,\sigma})}\\
&\leq 2\sqrt{\frac{L}{N}}\sqrt{\frac{8\min\{\mathrm{sp}(\widehat{V}^{\pi^\star,\sigma})_*,1/(1-\gamma))}{\gamma^2(1-\gamma)^2}}1\\
&\leq 8\sqrt{\frac{L}{\gamma^3(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}N}}1, \qquad (83)
\end{aligned}$$

where the last inequality is due to Lemma 5 for sa-rectangular case and with the same quantity replacing $\max\{1-\gamma,\sigma\}$ by $\max\{1-\gamma,\min_s\|\pi_s^*\|_*\tilde{\sigma}\}$ in the $s-$ rectangular case.

- For bounding $\mathcal{R}_5$, we can simply use (65)) to get

$$\mathcal{R}_5 = 2\sqrt{\frac{L}{N}}\Big(I-\gamma\underline{\widehat{P}}^{\pi^\star,\widehat{V}}\Big)^{-1}\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,\widehat{V}}}(V^{\pi^\star,\sigma}-\widehat{V}^{\pi^\star,\sigma})}$$
$$\leq 2\sqrt{\frac{L}{(1-\gamma)^2N}}\left\|V^{\star,\sigma}-\widehat{V}^{\pi^\star,\sigma}\right\|_\infty 1. \tag{84}$$

moreover,

$$\left\|V^{\star,\sigma}-\widehat{V}^{\pi^\star,\sigma}\right\|_\infty \leq \left\|V^{\star,\sigma}-\widehat{V}^{\pi^\star,\sigma}\right\|_\infty \leq \left\|V^{\star,\sigma}-\widehat{V}^{\pi^\star,\sigma}\right\|_\infty \tag{85}$$

as for $a>0, b>0$, we have $[a]-[b]<[a-b]$. Finally, we obtain

$$\mathcal{R}_5 \leq 2\sqrt{\frac{L}{(1-\gamma)^2N}}\left\|V^{\star,\sigma}-\widehat{V}^{\pi^\star,\sigma}\right\|_\infty 1. \tag{86}$$

- The term $\mathcal{R}_6$ can upper bounded as (78) as follows:

$$\mathcal{R}_6 \leq 2\sqrt{\frac{2L}{\gamma^2(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}N}}1. \tag{87}$$

for sa-rectangular case and with the same quantity replacing $\max\{1-\gamma,C_g\sigma\}$ by $\max\{1-\gamma,\min_s\|\pi_s^*\|_*\tilde{\sigma}C_g\}$ in the $s-$ rectangular case.

- Finally, $\mathcal{R}_7$ can be controlled the same as (79) shown below:

$$\mathcal{R}_7 \leq \frac{4L}{(1-\gamma)^2N}1. \tag{88}$$

Combining the results in (83), (86), (87), and (88) and inserting back to (82) leads to for $N \geq \frac{L}{(1-\gamma)^2}$

$$\Big(I-\gamma\underline{\widehat{P}}^{\pi^\star,\widehat{V}}\Big)^{-1}\Big(\underline{\widehat{P}}^{\pi^\star,V}V^{\pi^\star,\sigma}-\underline{P}^{\pi^\star,V}V^{\pi^\star,\sigma}\Big) \leq 8\sqrt{\frac{L(1+\frac{C_S\|1\|_*}{N(1-\gamma)})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}N}}1$$
$$+2\sqrt{\frac{L}{(1-\gamma)^2N}}\left\|V^{\star,\sigma}-\widehat{V}^{\pi^\star,\sigma}\right\|_\infty 1+2\sqrt{\frac{2L}{\gamma^2(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}N}}1+\frac{7LC_S\|1\|_*}{N(1-\gamma)^2}$$
$$\leq 80\sqrt{\frac{L(1+\frac{C_S\|1\|_*}{N(1-\gamma)})}{(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}N}}1+2\sqrt{\frac{L}{(1-\gamma)^2N}}\left\|V^{\star,\sigma}-\widehat{V}^{\pi^\star,\sigma}\right\|_\infty 1+\frac{7LC_S\|1\|_*}{N(1-\gamma)^2}, \tag{89}$$

where the last inequality follows from the assumption $\gamma \geq \frac{1}{4}$. Finally, inserting (81) and (89) back to (62) yields

$$\left\|\widehat{V}^{\pi^\star,\sigma}-V^{\pi^\star,\sigma}\right\|_\infty \leq \max\Bigg\{160\sqrt{\frac{L(1+\frac{C_S\|1\|_*}{N(1-\gamma)})}{(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}N}}+\frac{7LC_S\|1\|_*}{N(1-\gamma)^2},$$
$$80\sqrt{\frac{L(1+\frac{C_S\|1\|_*}{N(1-\gamma)})}{(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}N}}+2\sqrt{\frac{L}{(1-\gamma)^2N}}\left\|V^{\star,\sigma}-\widehat{V}^{\pi^\star,\sigma}\right\|_\infty+\frac{7LC_S\|1\|_*}{N(1-\gamma)^2}\Bigg\}$$
$$\leq 160\sqrt{\frac{L(1+\frac{C_S\|1\|_*}{N(1-\gamma)})}{(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}N}}+\frac{14LC_S\|1\|_*}{N(1-\gamma)^2}, \tag{90}$$

where the last inequality holds by taking $N \geq \frac{16\log(\frac{SAN}{\delta})}{(1-\gamma)^2}$ rearranging terms. In s-rectangular case, we obtain the same result, replacing $\max\{1-\gamma,C_g\sigma\}$ by $\max\{1-\gamma,\min_s\|\pi_s^*\|_*C_g\tilde{\sigma}\}$.

Third step: controlling $\|\widehat{V}^{\widehat{\pi},\sigma}-V^{\widehat{\pi},\sigma}\|_\infty$ or bounding the first and second term in (61)**.** Unlike the earlier term, one has to face a more complicated statistical dependency between $\widehat{\pi}$ and the empirical RMDP. To begin with, we introduce the following lemma which controls the main term on the right-hand side of (61), which is proved in Appendix 9.3.5.

Lemma 10. *Consider any $\delta\in(0,1)$. Taking $N\geq L''$ with probability at least $1-\delta$, one has for sa- or s-rectangular case :*

$$\left|\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}-\underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right|\leq 2\sqrt{\frac{L'}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma})}1+2\varepsilon_{\mathsf{opt}}1+\frac{15L''C_S\|1\|_*}{N(1-\gamma)}$$
$$\leq 2\sqrt{\frac{L''}{(1-\gamma)^2N}}1+2\varepsilon_{\mathsf{opt}}1+\frac{14L''C_S\|1\|_*}{N(1-\gamma)}1. \tag{91}$$

with $L''=2\log(\frac{54\|1\|_*SAN^2}{(1-\gamma)\delta})$. Moreover,For TV this lemma holds but without the geometric term $\frac{14L''C_S\|1\|_*}{N(1-\gamma)}1$. Taking the sup over s gives the final result.

With Lemma 10 in hand, we have to control **first term** in (61)

$$\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\left(\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}-\underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right)$$
$$\overset{\text{(i)}}{\leq}\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\left|\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}-\underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right|$$
$$\leq 2\sqrt{\frac{L'}{N}}\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{P^{\widehat{\pi}}}(\widehat{V}^{\star,\sigma})}+\left(I-\gamma\underline{P}^{\widehat{\pi},V^{\widehat{\pi}}}\right)^{-1}\left(2\varepsilon_{\mathsf{opt}}\right)1 \tag{92}$$
$$+\left(I-\gamma\underline{P}^{\widehat{\pi},V^{\widehat{\pi}}}\right)^{-1}\frac{14L''C_S\|1\|_*}{N(1-\gamma)}1$$
$$\overset{\text{(ii)}}{\leq}\left(\frac{2\varepsilon_{\mathsf{opt}}}{1-\gamma}\right)1+\underbrace{2\sqrt{\frac{L'}{N}}\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})}}_{=:\mathcal{S}_1}$$
$$+\underbrace{2\sqrt{\frac{L'}{N}}\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\left|\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\star,\sigma})-\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})\right|}}_{=:\mathcal{S}_2}$$
$$+\underbrace{2\sqrt{\frac{L'}{N}}\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\left|\mathrm{Var}_{P^{\widehat{\pi}}}(\widehat{V}^{\star,\sigma})-\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\star,\sigma})\right|}}_{=:\mathcal{S}_3}, \tag{93}$$

where (i) and (ii) hold by the fact that each row of $(1-\gamma)\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}$ is a probability vector that falls into $\Delta(\mathcal{S})$. The remainder of the proof will focus on controlling the three terms in (93) separately.

- For $\mathcal{S}_1$, we introduce the following lemma, whose proof is postponed to 9.3.6.

  Lemma 11. *Consider any $\delta\in(0,1)$. Taking $N\geq\frac{L''}{(1-\gamma)^2}$ one has with probability at least $1-\delta$, for sa$-$ rectangular*

$$\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})}\leq 6\sqrt{\frac{\left(1+\varepsilon_{\mathsf{opt}}+\frac{L''C_S\|1\|_*}{N(1-\gamma)}\right)}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}}}1$$
$$\leq 6\sqrt{\frac{\left(1+\varepsilon_{\mathsf{opt}}+\frac{L''C_S\|1\|_*}{N(1-\gamma)}\right)}{(1-\gamma)^3\gamma^3}}1.$$

  and for s-rectangular

$$
\begin{aligned}
\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})} &\le 6\sqrt{\frac{L''\left(1+\varepsilon_{\mathsf{opt}}+\frac{C_S\|1\|_*}{N(1-\gamma)}\right)}{\gamma^3(1-\gamma)^2\max\{1-\gamma, C_g\tilde{\sigma}\min_s\|\widehat{\pi}_s\|_\infty\}}}1 \\
&\le 6\sqrt{\frac{L''\left(1+\varepsilon_{\mathsf{opt}}+\frac{C_S\|1\|_*}{N(1-\gamma)}\right)}{(1-\gamma)^3\gamma^2}}1.
\end{aligned}
$$

Applying Lemma 11 and (65) to (93) leads to

$$
\begin{aligned}
\mathcal{S}_1 &= 2\sqrt{\frac{L'}{N}}\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})} \\
&\le 12\sqrt{\frac{L''}{\gamma^3(1-\gamma)^2\max\{1-\gamma, C_g\sigma\}N}}1.
\end{aligned} \tag{94}
$$

for sa-rectangular and the same quantity replacing $\max\{1-\gamma, C_g\sigma\}$ by $\max\{1-\gamma, C_g\tilde{\sigma}\min_s\|\widehat{\pi}_s\|_*\}$ for $s-$ rectangular case.

- Applying Lemma 1 with $\|\widehat{V}^{\star,\sigma}-\widehat{V}^{\widehat{\pi},\sigma}\|_\infty \le \varepsilon_{\mathsf{opt}}$ and (65), $\mathcal{S}_2$ can be controlled as

$$
\begin{aligned}
\mathcal{S}_2 &= 2\sqrt{\frac{L''}{N}}\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\left|\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\star,\sigma})-\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})\right|} \\
&\le 4\sqrt{\frac{L''}{N}}\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\varepsilon_{\mathsf{opt}}\frac{1}{1-\gamma}^2} \le 8\sqrt{\frac{\varepsilon_{\mathsf{opt}}L''}{(1-\gamma)^4N}}1.
\end{aligned} \tag{95}
$$

- $\mathcal{S}_3$ can be controlled similar to $\mathcal{R}_2$ in (78) as follows:

$$
\begin{aligned}
\mathcal{S}_3 &= 2\sqrt{\frac{L''}{N}}\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\left|\mathrm{Var}_{P^{\widehat{\pi}}}(\widehat{V}^{\star,\sigma})-\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\star,\sigma})\right|} \\
&\le 4\sqrt{\frac{L''}{N}}\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\frac{1}{\gamma^2\max\{1-\gamma, C_g\sigma\}}}1 \le 8\sqrt{\frac{L''}{\gamma^2(1-\gamma)^2\max\{1-\gamma, C_g\sigma\}N}}1
\end{aligned} \tag{96}
$$

for sa-rectangular and replacing $\max\{1-\gamma,\sigma\}$ by $\max\{1-\gamma,\tilde{\sigma}\min_s\|\widehat{\pi}_s\|_*\}$ for $s-$ rectangular case.

Finally, summing up the results in (94), (95), and (96) and inserting them back to (93) yields: taking $N \ge \frac{L''}{(1-\gamma)^2}$, with probability at least $1-\delta$,

$$
\begin{aligned}
&\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\left(\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}-\underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right) \le \left(\frac{2\varepsilon_{\mathsf{opt}}}{1-\gamma}\right)1+\frac{14L''C_S\|1\|_*}{N(1-\gamma)^2}1 \\
&+12\sqrt{\frac{L''\left(1+\varepsilon_{\mathsf{opt}}+\frac{C_S\|1\|_*}{N(1-\gamma)}\right)}{\gamma^3(1-\gamma)^2\max\{1-\gamma, C_g\sigma\}N}}1+8\sqrt{\frac{\varepsilon_{\mathsf{opt}}L'}{(1-\gamma)^4N}}1+8\sqrt{\frac{L'}{\gamma^2(1-\gamma)^2\max\{1-\gamma, C_g\sigma\}N}}1 \\
&\le 16\sqrt{\frac{L''\left(1+\varepsilon_{\mathsf{opt}}+\frac{C_S\|1\|_*}{N(1-\gamma)}\right)}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1+\left(\frac{2\varepsilon_{\mathsf{opt}}\gamma}{(1-\gamma)}+8\sqrt{\frac{\varepsilon_{\mathsf{opt}}\gamma L'}{(1-\gamma)^4N}}1+\frac{15L''C_S\|1\|_*}{N(1-\gamma)^2}1\right) \qquad (97)
\end{aligned}
$$

$$(98)$$

for sa-rectangular and the same quantity replacing $\max\{1-\gamma,\sigma\}$ by $\max\{1-\gamma,\tilde{\sigma}\min_s\|\widehat{\pi}_s\|_*\}$ for $s-$ rectangular case. In this step, it is harder to decouple terms as $\widehat{V}^{\widehat{\pi}}$ depends on data both in $\widehat{\pi}$ and $\widehat{V}$.

Step 5: controlling $\|\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\|_\infty$: bounding the second term in (61)**.** Towards this, applying Lemma 10 leads to in sa-rectangular case:

$$
\begin{aligned}
&\Big(I-\gamma \underline{P}^{\widehat{\pi},V}\Big)^{-1}\Big(\widehat{\underline{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\Big) \le \Big(I-\gamma \underline{P}^{\widehat{\pi},V}\Big)^{-1}\Big|\widehat{\underline{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\Big| \\
&\le 2\sqrt{\frac{L''}{N}}\Big(I-\gamma \underline{P}^{\widehat{\pi},V}\Big)^{-1}\sqrt{\mathrm{Var}_{P^{\widehat{\pi}}}(\widehat{V}^{\star,\sigma})} + \Big(I-\gamma \underline{P}^{\widehat{\pi},V}\Big)^{-1}\Bigg(2\varepsilon_{\mathsf{opt}}\Bigg)1 \qquad (99) \\
&+ \Big(I-\gamma \underline{P}^{\widehat{\pi},V}\Big)^{-1}\frac{L''14C_S\,\|1\|_*}{N(1-\gamma)}1 \\
&\le \Bigg(\frac{2\varepsilon_{\mathsf{opt}}}{(1-\gamma)}\Bigg)1 + \underbrace{2\sqrt{\frac{L''}{N}}\Big(I-\gamma \underline{P}^{\widehat{\pi},V}\Big)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},V}}(V^{\widehat{\pi},\sigma})}}_{=:\mathcal{S}_4} + \Big(I-\gamma \underline{P}^{\widehat{\pi},V}\Big)^{-1}\frac{L''C_S\,\|1\|_*}{N(1-\gamma)}1 \\
&\quad + \underbrace{2\sqrt{\frac{L'}{N}}\Big(I-\gamma \underline{P}^{\widehat{\pi},V}\Big)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},V}}(\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma})}}_{=:\mathcal{S}_5} \\
&\quad + \underbrace{2\sqrt{\frac{L''}{N}}\Big(I-\gamma \underline{P}^{\widehat{\pi},\widehat{V}}\Big)^{-1}\sqrt{\Big|\mathrm{Var}_{\underline{P}^{\widehat{\pi},V}}(\widehat{V}^{\star,\sigma}) - \mathrm{Var}_{\underline{P}^{\widehat{\pi},V}}([\widehat{V}^{\widehat{\pi},\sigma})\Big|}}_{=:\mathcal{S}_6} \\
&\quad + \underbrace{2\sqrt{\frac{L''}{N}}\Big(I-\gamma \underline{P}^{\widehat{\pi},\widehat{V}}\Big)^{-1}\sqrt{\Big|\mathrm{Var}_{P^{\widehat{\pi}}}(\widehat{V}^{\star,\sigma}) - \mathrm{Var}_{\underline{P}^{\widehat{\pi},V}}([\widehat{V}^{\star,\sigma})\Big|}}_{=:\mathcal{S}_7}. \qquad (100)
\end{aligned}
$$

We shall bound each of the terms separately.

- Applying Lemma 7 with $P = \underline{P}^{\widehat{\pi},V}$, $\pi = \widehat{\pi}$, and taking $V = V^{\widehat{\pi},\sigma}$ which obeys $V^{\widehat{\pi},\sigma} = r_{\widehat{\pi}} + \gamma \underline{P}^{\widehat{\pi},V}V^{\widehat{\pi},\sigma}$, the term $\mathcal{S}_4$ can be controlled similar to (83) as follows:

$$
\mathcal{S}_4 \le 8\sqrt{\frac{L''\Big(1+\varepsilon_{\mathsf{opt}} + \frac{C_S\|1\|_*}{N(1-\gamma)}\Big)}{\gamma^3(1-\gamma)^2\max\{1-\gamma, C_g\sigma\}N}}1. \qquad (101)
$$

  for sa-rectangular and the same quantity replacing $\max\{1-\gamma, C_g\sigma\}$ by $\max\{1-\gamma, \min_s\|\hat{\pi}_s\|_*\tilde{\sigma}C_g\}$ for $s-$ rectangular case.

- For $\mathcal{S}_5$, it is observed that

$$
\begin{aligned}
\mathcal{S}_5 &= 2\sqrt{\frac{L''}{N}}\Big(I-\gamma \underline{P}^{\widehat{\pi},V}\Big)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},V}}(\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma})} \\
&\le 2\sqrt{\frac{L''}{(1-\gamma)^2N}}\Big\|V^{\widehat{\pi},\sigma} - \widehat{V}^{\widehat{\pi},\sigma}\Big\|_\infty 1. \qquad (102)
\end{aligned}
$$

- Next, observing that $\mathcal{S}_6$ and $\mathcal{S}_7$ are almost the same as the terms $\mathcal{S}_2$ (controlled in (95)) and $\mathcal{S}_3$ (controlled in (96)) in (93), it is easily verified that they can be controlled as follows

$$
\mathcal{S}_6 \le 4\sqrt{\frac{\varepsilon_{\mathsf{opt}}L''}{(1-\gamma)^4N}}1, \qquad \mathcal{S}_7 \le 4\sqrt{\frac{L''}{\gamma^2(1-\gamma)^2\max\{1-\gamma, C_g\sigma\}N}}1. \qquad (103)
$$

for sa-rectangular and the same quantity replacing $\max\{1-\gamma,\sigma\}$ by $\max\{1-\gamma, \min_s \|\hat{\pi}_s\|_* \tilde{\sigma}\}$ for $s-$ rectangular case. Then inserting the results in (101), (102), and (103) back to (100) leads to

$$\left(I-\gamma \underline{P}^{\widehat{\pi},V}\right)^{-1}\left(\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}-\underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right) \tag{104}$$

$$\leq \left(\frac{2\varepsilon_{\mathsf{opt}}}{(1-\gamma)}\right)1+8\sqrt{\frac{L''\left(1+\varepsilon_{\mathsf{opt}}+\frac{C_S\|1\|_*}{N(1-\gamma)}\right)}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1+\frac{14L''C_S\|1\|_*}{N(1-\gamma)^2}1$$

$$+2\sqrt{\frac{L''}{(1-\gamma)^2N}}\left\|V^{\widehat{\pi},\sigma}-\widehat{V}^{\widehat{\pi},\sigma}\right\|_\infty 1+4\sqrt{\frac{L''\varepsilon_{\mathsf{opt}}}{(1-\gamma)^4N}}1+4\sqrt{\frac{L''}{\gamma^2(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}N}}1$$

$$\leq 12\sqrt{\frac{L''\left(1+\varepsilon_{\mathsf{opt}}+\frac{C_S\|1\|_*}{N(1-\gamma)}\right)}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}+4\sqrt{\frac{L''}{(1-\gamma)^2N}}\left\|V^{\widehat{\pi},\sigma}-\widehat{V}^{\widehat{\pi},\sigma}\right\|_\infty 1 \tag{105}$$

$$+\frac{3\varepsilon_{\mathsf{opt}}}{(1-\gamma)}+\frac{14L''C_S\|1\|_*}{N(1-\gamma)^2}1. \tag{106}$$

$$\tag{107}$$

Taking $N \geq \frac{16L''}{1-\gamma}$, we obtain $\frac{2\varepsilon_{\mathsf{opt}}}{(1-\gamma)}+4\varepsilon_{\mathsf{opt}}\sqrt{\frac{L''}{(1-\gamma)^4N}}1 \leq \frac{3\varepsilon_{\mathsf{opt}}}{(1-\gamma)}$ with probability at least $1-\delta$, inserting (97) and (105) back to (61)

$$\left\|\widehat{V}^{\widehat{\pi},\sigma}-V^{\widehat{\pi},\sigma}\right\|_\infty \leq \max\Big\{16\sqrt{\frac{L''\left(1+\varepsilon_{\mathsf{opt}}+\frac{C_S\|1\|_*}{N(1-\gamma)}\right)}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1+\left(\frac{2\varepsilon_{\mathsf{opt}}\gamma}{(1-\gamma)}+\frac{14L''C_S\|1\|_*}{N(1-\gamma)^2}1\right),$$

$$12\sqrt{\frac{L''\left(1+\varepsilon_{\mathsf{opt}}+\frac{C_S\|1\|_*}{N(1-\gamma)}\right)}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}+4\sqrt{\frac{L''}{(1-\gamma)^2N}}\left\|V^{\widehat{\pi},\sigma}-\widehat{V}^{\widehat{\pi},\sigma}\right\|_\infty 1 \tag{108}$$

$$+\frac{3\varepsilon_{\mathsf{opt}}}{(1-\gamma)}+\frac{14L''C_S\|1\|_*}{N(1-\gamma)^2}1.\Big\}$$

$$\leq 48\sqrt{\frac{L''\left(1+\varepsilon_{\mathsf{opt}}+\frac{C_S\|1\|_*}{N(1-\gamma)}\right)}{\gamma^3(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}N}}+\frac{6\varepsilon_{\mathsf{opt}}}{(1-\gamma)}+\frac{28L''C_S\|1\|_*}{N(1-\gamma)^2}1 \tag{109}$$

for sa-rectangular and the same quantity, replacing $\max\{1-\gamma,C_g\sigma\}$ by $\max\{1-\gamma,\tilde{\sigma}\min_s\|\hat{\pi}_s\|_*\}$ for $s-$ rectangular case. The proof is similar for TV without the geometric term depending on C_S.

Step 6: summing all the previous inequalities results. Using all the previous results in (90) and (109) and inserting back to (56) complete the proof as follows: taking $N \geq \frac{16L''}{(1-\gamma)^2}, \gamma > 1/4,$, with probability at least $1-\delta$, for sa-rectangular

$$\begin{aligned}\left\|V^{\star,\sigma}-V^{\widehat{\pi},\sigma}\right\|_\infty &\leq \left\|V^{\pi^\star,\sigma}-\widehat{V}^{\pi^\star,\sigma}\right\|_\infty+\varepsilon_{\mathsf{opt}}+\left\|\widehat{V}^{\widehat{\pi},\sigma}-V^{\widehat{\pi},\sigma}\right\|_\infty\\
&\leq \varepsilon_{\mathsf{opt}}+48\sqrt{\frac{L''\left(1+\varepsilon_{\mathsf{opt}}+\frac{C_S\|1\|_*}{N(1-\gamma)}\right)}{\gamma^3(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}N}}+\frac{6\varepsilon_{\mathsf{opt}}}{(1-\gamma)}+\frac{28L''C_S\|1\|_*}{N(1-\gamma)^2}1\\
&\quad+160\sqrt{\frac{L(1+\frac{C_S\|1\|_*}{N(1-\gamma)})}{(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}N}}+\frac{14LC_S\|1\|_*}{N(1-\gamma)^2}\\
&\leq \frac{8\varepsilon_{\mathsf{opt}}}{1-\gamma}+\frac{42L''C_S\|1\|_*}{N(1-\gamma)^2}+1508\sqrt{\frac{L''(1+\frac{C_S\|1\|_*}{N(1-\gamma)})}{(1-\gamma)^2\max\{1-\gamma,C_g\sigma\}N}}\end{aligned} \tag{110}$$

where the last inequality holds by $\gamma \geq \frac{1}{4}$ and $N \geq \frac{16L''}{(1-\gamma)^2}$ for sa-rectangular and the same quantity replacing $\max\{1-\gamma, \sigma\}$ by $\max\{1-\gamma, \tilde{\sigma} \min_s\{\|\pi_s^*\|_*\}\}$ for $s-$ rectangular case. The proof is similar for TV without the geometric term depending on C_S.

9.3 Proof of the auxiliary lemmas

9.3.1 Proof of Lemma 5

Similarly to Shi et al. [2023], denoting s_0 the argmax of $V^{\pi,\sigma}$ such that $V^{\pi,\sigma}(s_0) = \min_{s\in\mathcal{S}} V^{\pi,\sigma}(s)$ using recursive Bellman's equation

$$\max_{s\in\mathcal{S}} V^{\pi,\sigma}(s) = \max_{s\in\mathcal{S}} \mathbb{E}_{a\sim\pi(\cdot|s)} \left[r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{s,a})} \mathcal{P}V^{\pi,\sigma} \right] \tag{111}$$

$$\leq \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(1 + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{s,a})} \mathcal{P}V^{\pi,\sigma} \right) \tag{112}$$

where the second line holds since the reward function $r(s,a) \in [0,1]$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$.

Then we construct for any $(s,a) \in \mathcal{S}\times\mathcal{A}$ $\widetilde{P}_{s,a} \in \mathbb{R}^S$ by reducing the values of some elements of $P_{s,a}$ such that $P_{s,a} \geq \widetilde{P}_{s,a} \geq 0$ and $\sum_{s'} \left(P_{s,a}(s') - \widetilde{P}_{s,a}(s')\right) = \sigma C_g^{s,a}$. with $C_g^{s,a} = \frac{1}{\|e_{s_0}\|}$ It lead to $\widetilde{P}_{s,a} + \sigma C_g^{s,a} e_{s_0}^\top \in \mathcal{U}^\sigma_{\|\|}(P_{s,a})$, where e_{s_0} is the standard basis vector supported on s_0, since

$$\frac{1}{2}\left\| \widetilde{P}_{s,a} + \sigma C_g^{s,a} e_{s_0}^\top - P_{s,a} \right\| \leq \frac{1}{2}\left\| \widetilde{P}_{s,a} - P_{s,a} \right\| + \frac{C_g^{s,a}\sigma \|e_{s_0}\|}{2} = \sigma/2 + \sigma/2 = \sigma \tag{113}$$

Consequently,

$$\inf_{\mathcal{P}\in\mathcal{U}^\sigma_{\|\cdot\|}(P_{s,a})} \mathcal{P}V^{\pi,\sigma} \leq \left(\widetilde{P}_{s,a} + \sigma C_g^{s,a} e_{s_0}^\top \right) V^{\pi,\sigma} \leq \left\| \widetilde{P}_{s,a} \right\|_1 \|V^{\pi,\sigma}\|_\infty + \sigma V^{\pi,\sigma}(s_0) C_g \tag{114}$$

$$\leq (1 - C_g^{s,a}\sigma) \max_{s\in\mathcal{S}} V^{\pi,\sigma}(s) + \sigma C_g^{s,a} \min_{s\in\mathcal{S}} V^{\pi,\sigma}(s) \tag{115}$$

where the second inequality holds by

$$\left\| \widetilde{P}_{s,a} \right\|_1 = \sum_{s'} \widetilde{P}_{s,a}(s') = -\sum_{s'} \left(P_{s,a}(s') - \widetilde{P}_{s,a}(s') \right) + \sum_{s'} P_{s,a}(s') = 1 - \sigma C_g^{s,a} \tag{116}$$

Plugging this back to the previous relation gives

$$\max_{s\in\mathcal{S}} V^{\pi,\sigma}(s) \leq 1 + \gamma(1 - C_g^{s,a}\sigma) \max_{s\in\mathcal{S}} V^{\pi,\sigma}(s) + \gamma C_g^{s,a}\sigma \min_{s\in\mathcal{S}} V^{\pi,\sigma}(s) \tag{117}$$

which, by rearranging terms, yields

$$\max_{s\in\mathcal{S}} V^{\pi,\sigma}(s) \leq \frac{1 + \gamma C_g^{s,a}\sigma \min_{s\in\mathcal{S}} V^{\pi,\sigma}(s)}{1 - \gamma(1 - C_g^{s,a}\sigma)} \tag{118}$$

$$\leq \frac{1}{(1-\gamma) + \gamma C_g^{s,a}\sigma} + \min_{s\in\mathcal{S}} V^{\pi,\sigma}(s) \leq \frac{1}{\gamma \max\{1-\gamma, C_g^{s,a}\sigma\}} + \min_{s\in\mathcal{S}} V^{\pi,\sigma}(s) \tag{119}$$

So rearranging term it holds :

$$\mathrm{sp}(V^{\pi,\sigma})_\infty \leq \frac{1}{\gamma \max\{1-\gamma, C_g\sigma\}} \tag{120}$$

As we pick the supreme over s ov this quantity, $C_g^{s,a}$ is replaced by $C_g = 1/(\min_s \|e_s\|)$ to obtain a control for every s.

9.3.2 Proof of Lemma 6

Similarly to 5 denoting s_0 the argmax of $V^{\pi,\sigma}$ such that $V^{\pi,\sigma}(s_0) = \min_{s\in\mathcal{S}} V^{\pi,\sigma}(s)$ using recursive Bellman's equation

$$\max_{s\in\mathcal{S}} V^{\pi,\sigma}(s) = \max_{s\in\mathcal{S}} \mathbb{E}_{a\sim\pi(\cdot|s)}\left[r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P_s)} \mathcal{P}V^{\pi,\tilde{\sigma}}\right] \tag{121}$$

$$\leq \max_{(s)\in\mathcal{S}}\left(1 + \gamma \inf_{\mathcal{P}^{\pi}\in\mathcal{U}^{\sigma}(P_s^{\pi})} \mathcal{P}^{\pi}V^{\pi,\tilde{\sigma}}\right) \tag{122}$$

where the second line holds since the reward function $r(s,a)\in[0,1]$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$.Then we construct for any $(s)\in\mathcal{S}$ $\widetilde{P}_s\in\mathbb{R}^{S\times A}$ by reducing the values of some elements of P_s such that $P_s \geq \widetilde{P}_s \geq 0$ and

$$\forall a\in A, \sum_{s'}\left(P_s(s',a) - \widetilde{P}_s(s',a)\right) = \sigma_{s,a}C_g^s$$

Writting $\|\sigma_{s,a}\| \leq \tilde{\sigma}$ we construction $\sigma_{s,a}$ such that

$$\sum_a \pi(a|s)\sum_{s'}\left(P_s(s',a) - \widetilde{P}_s(s',a)\right) = \|\pi_s\|_* \tilde{\sigma}C_g^s \tag{123}$$

Not that this construction is possible as it is simply Cauchy Swartz equality case.

It leads to $\widetilde{P}_s + \sigma e_{s_0,a}^{\top}\in\mathcal{U}^{\tilde{\sigma}}(P_s)$, where $e_{s_0,a}\in\mathbb{R}^{S\times A}$ is the standard basis vector supported on s_0 which is equal to 1 at s_0 for every a and otherwise.

$$\frac{1}{2}\left\|\widetilde{P}_s + \sigma_{s,a}C_g^s e_{s_0,a}^{\top} - P_s\right\| \leq \frac{1}{2}\left\|\widetilde{P}_s - P_s\right\| + \frac{\tilde{\sigma}\left\|e_{s_0}\right\|C_g}{2} = \tilde{\sigma}/2 + \tilde{\sigma}/2 \tag{124}$$

as $C_g^s\left\|\sigma_{s,a}e_{s_0,a}\right\|$ is equal to $C_g^s\tilde{\sigma}\left\|e_{s_0}\right\|$ Consequently,

$$\inf_{\mathcal{P}^{\pi}\in\mathcal{U}^{\sigma}(P_s)} \mathcal{P}^{\pi}V^{\pi,\tilde{\sigma}} \leq \Pi^{\pi}\left(\widetilde{P}_s^{\pi} + \sigma C_g^s e_{s_0}^{\top}\right)V^{\pi,\tilde{\sigma}} \tag{125}$$

$$= \sum_a\sum_{s'}\widetilde{P}_s(s',a)\pi(a|s)V^{\pi,\tilde{\sigma}}(s') + \sigma e_{s_0,a}C_g^s V^{\pi,\tilde{\sigma}}(s_0)\pi(a|s) \tag{126}$$

$$= \sum_a \sup_{s'} V(s')(\sum_{s'}\widetilde{P}_s(s',a)))\pi(a|s) + V^{\pi,\tilde{\sigma}}(s_0)\pi(a|s)\sigma_{s,a}C_g^s \tag{127}$$

$$\stackrel{(a)}{=} \max_{s\in\mathcal{S}} V^{\pi,\sigma}(s)\sum_a(1-\sigma C_g^s)\pi(a|s) + \sum_a V^{\pi,\tilde{\sigma}}(s_0)\pi(a|s)\sigma_{s,a}C_g^s \tag{128}$$

$$\stackrel{(b)}{=} \max_{s\in\mathcal{S}} V^{\pi,\sigma}(s)(1-\tilde{\sigma}C_g^s)\left\|\pi_s\right\|_* + \left\|\pi_s\right\|_*\tilde{\sigma}C_g^s \min_{s\in\mathcal{S}} V^{\pi,\tilde{\sigma}}(s) \tag{129}$$

$$\leq (1 - C_g^s\tilde{\sigma})\max_{s\in\mathcal{S}} V^{\pi,\sigma}(s) + \sigma C_g^s \min_{s\in\mathcal{S}} V^{\pi,\tilde{\sigma}}(s) \tag{130}$$

where $\|\pi\|_\infty$ is the norm of the vector $\pi(.|s)$ and where (a) holds because

$$\sum_{s'}\widetilde{P}_s(s') = -\sum_{s'}\left(P_s(s') - \widetilde{P}_s(s')\right) + \sum_{s'}P_s(s') = 1 - \sigma_{s,a}C_g^s \tag{131}$$

Finally (b) is due to (123). Plugging this back to the previous relation gives

$$\max_{s\in\mathcal{S}} V^{\pi,\tilde{\sigma}}(s) \leq 1 + \gamma(1 - \tilde{\sigma}C_g^s\left\|\pi_s\right\|_*)\max_{s\in\mathcal{S}} V^{\pi,\sigma}(s) + \gamma\left\|\pi_s\right\|_*\sigma C_g^s \min_{s\in\mathcal{S}} V^{\pi,\tilde{\sigma}}(s) \tag{132}$$

which, by rearranging terms, yields

$$
\begin{aligned}
\max_{s\in\mathcal{S}} V^{\pi,\tilde{\sigma}}(s) &\le \frac{1+\gamma\tilde{\sigma}\left\|\pi_s\right\|_* C_g^s \min_{s\in\mathcal{S}} V^{\pi,\tilde{\sigma}}(s)}{1-\gamma(1-C_g^s\tilde{\sigma}\left\|\pi_s\right\|_*)} && (133)\\
&\le \frac{1}{(1-\gamma)+\left\|\pi_s\right\|_*\gamma C_g^s\tilde{\sigma}} + \min_{s\in\mathcal{S}} V^{\pi,\tilde{\sigma}}(s) && (134)\\
&\le \frac{1}{(1-\gamma)+\gamma\left\|\pi_s\right\|_* C_g^s\tilde{\sigma}} + \min_{s\in\mathcal{S}} V^{\pi,\tilde{\sigma}}(s) && (135)\\
&\le \frac{1}{\gamma\max\{1-\gamma, C_g^s\left\|\pi_s\right\|_*\tilde{\sigma}\}} + \min_{s\in\mathcal{S}} V^{\pi,\tilde{\sigma}}(s) && (136)
\end{aligned}
$$

So rearranging and taking the sumpremum over all $s$term it holds :

$$
\mathrm{sp}(V^{\pi,\tilde{\sigma}})_\infty \le \frac{1}{\gamma\max\{1-\gamma, \min_s\left\|\pi_s\right\|_* C_g\tilde{\sigma}\}} \tag{137}
$$

As we pick the supreme over s ovf this quantity, C_g^s is replaced by $C_g = 1/\min_s\left\|e_s\right\|$

9.3.3 Proof of Lemma 8

Proof. Concentration of the robust values function. with probability $1-\delta$, it holds:

$$
\left|P_{s,a}^{\pi,V}V - \widehat{P}_{s,a}^{\pi,V}V\right| \le 2\sqrt{\frac{L}{N}}\sqrt{\mathrm{Var}[V]_{\alpha^{**}}}(V) + \frac{3LC_S\left\|1\right\|_*}{N(1-\gamma)}
$$

with $L = 2\log(18\left\|1\right\|_* SAN/\delta)$ and First we can use optimization duality such as in (50):

$$
\begin{aligned}
&\left|P_{s,a}^{\pi,V}V - \widehat{P}_{s,a}^{\pi,V}V\right| && (138)\\
&= \Big| \max_{\mu^{\lambda,\omega}_{P^0_{s,a}}\in\mathcal{M}^{\lambda,\omega}_{P^0_{s,a}}} \left\{P^0_{s,a}(V-\mu) - \sigma\left(\mathrm{sp}((V-\mu))_*\right)\right\}\\
&- \max_{\mu^{\lambda,\omega}_{\hat{P}^0_{s,a}}\in\mathcal{M}^{\lambda,\omega}_{\hat{P}^0_{s,a}}} \left\{\widehat{P}^0_{s,a}(V-\mu^{\lambda,\omega}_{\hat{P}^0_{s,a}}) - \sigma\left(\mathrm{sp}((V-\mu^{\lambda,\omega}_{\hat{P}^0_{s,a}}))_*\right)\right\}\Big|\\
&\le \max\Big\{\Big| \max_{\mu^{\lambda,\omega}_{P^0_{s,a}}\in\mathcal{M}^{\lambda,\omega}_{P^0_{s,a}}} \left\{P^0_{s,a}(V-\mu^{\lambda,\omega}_{P^0_{s,a}}) - \sigma\left(\mathrm{sp}((V-\mu^{\lambda,\omega}_{P^0_{s,a}}))_*\right)\right\}\\
&- \max_{\mu^{\lambda,\omega}_{P^0_{s,a}}\in\mathcal{M}^{\lambda,\omega}_{P^0_{s,a}}} \left\{\widehat{P}^0_{s,a}(V-\mu^{\lambda,\omega}_{P^0_{s,a}}) - \sigma\left(\mathrm{sp}((V-\mu^{\lambda,\omega}_{P^0_{s,a}}))_*\right)\right\}\Big|; && (139)\\
&\Big| \max_{\mu^{\lambda,\omega}_{\hat{P}^0_{s,a}}\in\mathcal{M}^{\lambda,\omega}_{\hat{P}^0_{s,a}}} \left\{\widehat{P}^0_{s,a}(V-\mu^{\lambda,\omega}_{\hat{P}^0_{s,a}}) - \sigma\left(\mathrm{sp}((V-\mu^{\lambda,\omega}_{\hat{P}^0_{s,a}}))_*\right)\right\} && (140)\\
&- \max_{\mu^{\lambda,\omega}_{\hat{P}^0_{s,a}}\in\mathcal{M}^{\lambda,\omega}_{\hat{P}^0_{s,a}}} \left\{P^0_{s,a}(V-\mu^{\lambda,\omega}_{\hat{P}^0_{s,a}}) - \sigma\left(\mathrm{sp}((V-\mu^{\lambda,\omega}_{\hat{P}^0_{s,a}}))_*\right)\right\} \quad \Big|\Big\}\\
&\le \max\Big\{\underbrace{\left|\max_{\mu\in\mu^{\lambda,\omega}_{P^0_{s,a}}}\left(P^0_{s,a}-\widehat{P}^0_{s,a}\right)(V-\mu^{\lambda,\omega}_{P^0_{s,a}})\right|}_{=:g_{s,a}(\alpha_P^{\lambda,\omega},V)}, \underbrace{\left|\max_{\mu^{\lambda,\omega}_{\hat{P}^0_{s,a}}\in\mathcal{M}^{\lambda,\omega}_{\hat{P}^0_{s,a}}}\left(P^0_{s,a}-\widehat{P}^0_{s,a}\right)(V-\mu^{\lambda,\omega}_{\hat{P}^0_{s,a}})\right|}_{=:g_{s,a}(\alpha_{\hat{P}}^{\lambda,\omega},V)}\Big\} && (141)
\end{aligned}
$$

where in the first equality we use Lemma 3. The final inequality is a consequence of the 1-Lipschitzness of the max operator. First, we control $g_{s,a}(\alpha_P^{\lambda,\omega},V)$. To do so, we use for a fixed $\alpha_P^{\lambda,\omega}$

and any vector V that is independent with $\widehat{P}^0$, the Bernstein's inequality, one has with probability at least $1-\delta$ with sa-rectangular notations,

$$g_{s,a}(\alpha_P^{\lambda,\omega}, V) = \left| \left(P^0_{s,a} - \widehat{P}^0_{s,a} \right) [V]_{\alpha_P^{\lambda,\omega}} \right| \leq \sqrt{\frac{2\log(\frac{2}{\delta})}{N}} \sqrt{\mathrm{Var}_{P^0_{s,a}}(V)} + \frac{2\log(\frac{2}{\delta})}{3N(1-\gamma)}. \tag{142}$$

Once pointwise concentration derived, we will use uniform concentration to yield this lemma. First, union bound, is obtained noticing that $g_{s,a}(\alpha_P^{\lambda,\omega}, V)$ is 1-Lipschitz w.r.t. λ and ω as it is linear in λ and ω. Moreover, $\lambda^* = \|V - \mu^* - \omega\|_*$ obeying $\lambda^* \leq \frac{\|1\|_*}{1-\gamma}$. The quantity $\omega \in [0, 1/(1-\gamma)]$ as it is always smaller that V by definition. We construct then a 2-dimensional a ε_1-net N_{ε_1} over $\lambda^* \in [0, \frac{\|1\|_*}{1-\gamma}]$ and $\omega \in [0, 1/(1-\gamma)]$ whose size satisfies $|N_{\varepsilon_1}| \leq \left(\frac{3\|1\|_*}{\varepsilon_1(1-\gamma)} \right)^2$ [Vershynin, 2018]. Using union bound and (142), it holds with probability at least $1 - \frac{\delta}{SA}$ that for all $\lambda \in N_{\varepsilon_1}$,

$$g_{s,a}(\alpha_P^{\lambda}, V) \leq \sqrt{\frac{2\log(\frac{2SA|N_{\varepsilon_1}|}{\delta})}{N}} \sqrt{\mathrm{Var}_{P^0_{s,a}}(V)} + \frac{2\log(\frac{2SA|N_{\varepsilon_1}|}{\delta})}{3N(1-\gamma)}. \tag{143}$$

Using the previous equation and also (141), it results in using notation $2\log(\frac{18SAN}{\delta}) = L$,

$$\begin{aligned}
g_{s,a}(\alpha_P^{\lambda}, V) &\overset{(a)}{\leq} \sup_{\alpha_P^{\lambda} \in N_{\varepsilon_1}} \left| \left(P^0_{s,a} - \widehat{P}^0_{s,a} \right) [V]_{\alpha_P^{\lambda}} \right| + \varepsilon_1 \\
&\overset{(b)}{\leq} \sqrt{\frac{2\log(\frac{2SA|N_{\varepsilon_1}|}{\delta})}{N}} \sqrt{\mathrm{Var}_{P^0_{s,a}}(V)} + \frac{2\log(\frac{2SA|N_{\varepsilon_1}|}{\delta})}{3N(1-\gamma)} + \varepsilon_1 \qquad (144) \\
&\overset{(c)}{\leq} \sqrt{\frac{2\log(\frac{2SA|N_{\varepsilon_1}|}{\delta})}{N}} \sqrt{\mathrm{Var}_{P^0_{s,a}}(V)} + \frac{\log(\frac{2SA|N_{\varepsilon_1}|}{\delta})}{N(1-\gamma)} \\
&\overset{(d)}{\leq} 2\sqrt{\frac{L}{N}} \sqrt{\mathrm{Var}_{P^0_{s,a}}(V)} + \frac{L}{N(1-\gamma)} \qquad (145) \\
&\leq 2\sqrt{\frac{L}{N}} \|V\|_\infty + \frac{L}{N(1-\gamma)} \\
&\leq 3\sqrt{\frac{L}{(1-\gamma)^2 N}} \qquad (146)
\end{aligned}$$

where (a) is because the optimal α^* falls into the ε_1-ball centered around some point inside N_{ε_1} and $g_{s,a}(\alpha_P^{\lambda}, V)$ is 1-Lipschitz with regard to λ and ω, (b) is due to Eq. (143), (c) arises from taking $\varepsilon_1 = \frac{\log(\frac{2SA|N_{\varepsilon_1}|}{\delta})}{3N(1-\gamma)}$, (d) is verified by $|N_{\varepsilon_1}| \leq \left(\frac{3\|1\|_*}{\varepsilon_1(1-\gamma)} \right)^2 \leq 9N\|1\|$ and that variance of a ceiling function of a vector is smaller than the variance of non-ceiling vector , and the last inequality comes from the fact $\|V^{\star,\sigma}\|_\infty \leq \frac{1}{1-\gamma}$ and taking $N \geq 2\log(\frac{18SAN\|1\|_*}{\delta}) = L$.

Contrary to the previous term, the second term $g_{s,a}(\alpha_{\hat{P}}^{\lambda}, V)$ is more difficult as we need concentration, but there is an extra dependency in the data thought the parameter $\alpha_{\hat{P}}^{\lambda}$. We need to decouple this problem using absorbing MDPs. Then it leads to

$$g_{s,a}(\alpha^{\lambda,\omega}_{\hat{P}}, V) \tag{147}$$

$$= | \max_{\mu^{\lambda,\omega}_{\hat{P}^0_{s,a}} \in \mathcal{M}^{\lambda,\omega}_{\hat{P}^0_{s,a}}} \left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)(V - \mu^{\lambda,\omega}_{\hat{P}^0_{s,a}})| \tag{148}$$

$$= | \max_{\mu \in \mathcal{M}^{\lambda,\omega}_{\hat{P}^0_{s,a}}} \left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)(V - \mu^{\lambda,\omega}_{P^0_{s,a}}) + \left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)(\mu^{\lambda,\omega}_{P^0_{s,a}} - \mu^{\lambda,\omega}_{\hat{P}^0_{s,a}})| \tag{149}$$

$$\leq | \max_{\mu^{\lambda,\omega}_{P^0_{s,a}} \in \mathcal{M}^{\lambda,\omega}_{P^0_{s,a}}} \left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)(V - \mu^{\lambda,\omega}_{P^0_{s,a}}) + \max_{\mu^{\lambda,\omega}_{\hat{P}^0_{s,a}} \in \mathcal{M}^{\lambda,\omega}_{\hat{P}^0_{s,a}}} \left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)(\mu^{\lambda,\omega}_{P^0_{s,a}} - \mu^{\lambda,\omega}_{\hat{P}^0_{s,a}})| \tag{150}$$

In the first equality, we add the term $\mu^{\lambda,\omega}_{P^0_{s,a}}$ to retrieve the previous concentration problem, fixing $P^0_{s,a}$ and optimizing λ, ω. In the second, we extend the max using triangular inequality. The first term in the last equality is exactly the term we have controlled previously, while the second one needs more attention. We decouple the dependency of the data, and then controlling the difference between the μ. Then using the characterization of the optimal μ from equation (47):

$$\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)(\mu^{\lambda,\omega}_{P^0_{s,a}} - \mu^{\lambda,\omega}_{\hat{P}^0_{s,a}}) = \sum_{s'} \lambda \left(P^0_{s,a}(s') - \widehat{P}^0_{s,a}(s')\right)(\nabla \left\|P^0_{s,a}\right\| - \nabla \left\|\hat{P}^0_{s,a}\right\|)$$

Here we assume that the subgradient are gradient as we assume that the norm is C^2. The question that arises is whether the gradient if the norm is Lipschitz. Assuming that the norm is C^2, using Mean value theorem, we know that

$$\left\|(\nabla \left\|P^0_{s,a}\right\| - \nabla \left\|\hat{P}^0_{s,a}\right\|)\right\|_2 \leq \sup_{x \in \Delta(S)} \left\|\nabla^2 \left\|x\right\|\right\|_2 \left\|(P^0_{s,a} - \hat{P}^0_{s,a})\right\|_2.$$

As the norm is C^2, is continuous and as the simplex is bounded, this quantity exists according to Extreme value theorem. It is possible to compute this contact depending on S for explicit norm such as L_p. Indeed, for L_2:

$$\nabla^2 \left\|x\right\|_2 = \frac{(I - \frac{x \bigotimes x)}{\|x\|_2^2}}{\left\|x\right\|_2} \leq \frac{1}{\left\|x\right\|_2} I \leq \frac{1}{\min_{x \in \Delta(S)} \left\|x\right\|_2} I = \sqrt{S}$$

where $\bigotimes$ is the Kronecker product. So we have an upper bound independently of x. For $L_p = \left\|x\right\|_p$ norms, $p \geq 2$, we have simple taking derivative twice:

$$\nabla^2 \left\|x\right\|_p = \frac{p-1}{L_p}\left(\mathcal{A}^{p-2} - g_p g_p^T\right)$$

with

$$\mathcal{A} = \mathrm{Diag}\left(\frac{\mathrm{abs}(x)}{L_p}\right)$$
$$g_p = \mathcal{A}^{p-2}\left(\frac{x}{L_p}\right).$$

where Diag is the diagonal matrix. However, as $x \leq L_p$, $\mathcal{A} \leq I$, we get

$$H \leq \frac{p-1}{\left\|x\right\|_p} \leq (p-1) S^{1/q} = C_S \tag{151}$$

where the $1/L_p$ is minimized for the uniform distribution. Then using Cauchy Swartz inequality, it holds

$$\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)(\mu^{\lambda,\omega}_{P^0_{s,a}} - \mu^{\lambda,\omega}_{\hat{P}^0_{s,a}}) \leq \lambda \left\|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\right\|_2^2. \tag{152}$$

Then the question is how to bound the quantity $\left\|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\right\|_2^2$. To do so, we will use Mac Diarmid inequality.

Definition 3. *Bounded difference property*

A function $f : \mathcal{X}_1 \times \ldots \mathcal{X}_n \to \mathbb{R}$ satisfies the bounded difference property if for each $i = 1, \ldots, n$ the change of coordinate from s_i to s_i' may change the value of the function at most on c_i

$$\forall i \in [n] : \sup_{x_i' \in \mathcal{X}_i} |f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i$$

In our case, we consider $f(X_1, \ldots, X_n) = \|\sum_{k=1}^n X_k\|_2$. Then we can notice that by triangle inequality for any $x_1, \ldots, x_n$ and x_k' with $X_{i,s'} = P^0_{i,s,a}(s') - P^0_{s,a}(s')$ (index i holds for index of sample generated from the generative model) that

$$\begin{aligned}
&f(x_1, \ldots, x_k, \ldots, x_n) = \|x_1 + \ldots + x_n\|_2 \leq \|x_1 + \ldots + x_n - x_k + x_k'\|_2 + \|x_k - x_k'\|_2 \\
&\leq f(x_1, \ldots, x_k', \ldots, x_n) + 2
\end{aligned}$$

Theorem 5. *(McDiarmid's inequality). McDiarmid et al. [1989] Let $f : \mathcal{X}_1 \times \ldots \mathcal{X}_n \to \mathbb{R}$ be a function satisfying the bounded difference property with bounds $c_1, \ldots, c_n$. Consider independent random variables $X_1, \ldots, X_n, X_i \in \mathcal{X}_i$ for all i. Then for any $t > 0$*

$$\mathbb{P}[f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] \geq t] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right)$$

Using McDiarmid's inequality and union bound, we can bound the term as here

$$\left\|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\right\|_2^2 - \mathbb{E}[\left\|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\right\|_2^2] \leq \frac{2N\log(|S||A|/\delta))}{N^2}$$

with probability $1 - \delta/(|S||A|)$. Moreover, the additional term can be bounded as follows:

$$\mathbb{E}[\left\|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\right\|_2^2] = \mathbb{E}[\sum_{s'} (P^0_{s,a}(s') - P^0_{s,a}(s'))^2 = \mathbb{E}[\sum_{s'} (\frac{1}{N}\sum_i^N X_{i,s'})^2]$$

with $X_{i,s'} = P^0_{i,s,a}(s') - P^0_{s,a}(s')$ is one sample sampled from the generative model. Then

$$\begin{aligned}
\mathbb{E}[\left\|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\right\|_2^2] = \frac{1}{N^2}\sum_{s'} \mathsf{Var}(\sum_i^N X_{i,s}) &\overset{a}{=} \frac{1}{N^2}\sum_i^N \sum_{s'} \mathsf{Var}(X_{i,s}) \\
&= \frac{1}{N^2}\sum_i^N \mathbb{E}(\sum_{s'} X_{i,s}^2) \leq \frac{4}{N}
\end{aligned}$$

where (a) the last equality comes from the independence of the random variables and where the last inequality comes from the fact the maximum of two elements in the simplex is bounded by 2. Finally, regrouping the two terms, we obtain with probability $1 - \delta/(|S||A|)$:

$$\begin{aligned}
\left\|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\right\|_2^2 &\leq \frac{2N\log(|S||A|/(\delta)))}{N^2} + \frac{4}{N} = \frac{8\log(|S||A|/(\delta)))}{N} + \frac{4}{N} \\
&\leq \frac{6\log(|S||A|/(\delta))}{N} = \frac{L'}{N}
\end{aligned}$$

with $L' = 6\log(|S||A|/(\delta))$. Finally, plugging the previous equation in (152):

$$\max_{\mu \in \mu^\lambda_{\widehat{P}^0_{s,a}}} \left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)(\mu^\lambda_{P^0_{s,a}} - \mu)| \leq \max_\lambda \left\|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\right\|_2^2 C_S \lambda.$$

This term can be easily controlled by taking the supremum over λ which is a 1 dimensional parameter. Then we can bound $\lambda \in [0, H\,\|1\|_*]$. Indeed,

$$\lambda^* = \|V - \mu^* - \eta\|_* \leq \|V\|_* \leq H\,\|1\|_* .$$

Finally, we obtain:

$$\max_{\lambda} \left\| \left(P^0_{s,a} - \widehat{P}^0_{s,a} \right) \right\|_2^2 C_S \lambda \leq \frac{L' C_S \|1\|_*}{N(1-\gamma)}.$$

Regrouping all terms:

$$\begin{aligned} g_{s,a}(\alpha^{\lambda}_{\hat{P}}, V) &\leq | \max_{\mu^{\lambda}_{P^0_{s,a}} \in \mathcal{M}^{\lambda}_{P^0_{s,a}}} \left(P^0_{s,a} - \widehat{P}^0_{s,a} \right) (V - \mu^{\lambda}_{P^0_{s,a}}) + \max_{\mu^{\lambda}_{\hat{P}^0_{s,a}} \in \mathcal{M}^{\lambda}_{\hat{P}^0_{s,a}}} \left(P^0_{s,a} - \widehat{P}^0_{s,a} \right) (\mu^{\lambda}_{P^0_{s,a}} - \mu^{\lambda}_{\hat{P}^0_{s,a}}) | \\ &\leq 2\sqrt{\frac{L}{N}} \sqrt{\mathrm{Var}(V)} + \frac{L' C_S \|1\|_*}{N(1-\gamma)} + \frac{L}{N(1-\gamma)} \leq 2\sqrt{\frac{L}{N}} \sqrt{\mathrm{Var}(V)} + \frac{3 L C_S \|1\|_*}{N(1-\gamma)} \end{aligned} \tag{153}$$

$$\tag{154}$$

We can recognize that the second term is a second order term as long as $N \geq (C_S \|1\|_*)^2$, we can regroup the two terms. Finally, as $g_{s,a}(\alpha^{\lambda}_{\hat{P}}, V) \geq g_{s,a}(\alpha^{\lambda}_{P}, V)$, we obtain

$$\left| P^{\pi,V}_{s,a} V - \widehat{P}^{\pi,V}_{s,a} V \right| \leq 2\sqrt{\frac{L}{N}} \sqrt{\mathrm{Var}_{P^0_{s,a}}(V)} + \frac{3 L C_S \|1\|_*}{N(1-\gamma)} \tag{155}$$

It is important to note that the geometry of the norm is present in the second order term $\frac{3LC_S\|1\|}{N(1-\gamma)}$ but this term is negligible as it is proportional to $1/N$ with regard to the variance term in $1/\sqrt{N}$. Moreover, note that the quantity $C_S \|1\|_* = S$ for L_2 norms.

For the specific case of TV which is not C^2 smooth, this lemma still holds as in (141), we only need to control one term without the dependency on data in the supremum as α^{λ}_{P} reduces to a scalar α which does not depend on P. Then extra decomposition using smoothness of the norm is not needed, as the only remaining term in the max in (141) is the left hand side term.

For the s-rectangular case, the first equation can be rewritten simply factorizing by $\pi(a|s)$ using lemma 4.

$$\begin{aligned} \left| P^{\pi,V}_{s,a} V - \widehat{P}^{\pi,V}_{s,a} V \right| &= \Big| \sum_a \pi(a|s) \max_{\mu^{\lambda}_{P^0_{s,a}} \in \mathcal{M}^{\lambda}_{P^0_{s,a}}} \left\{ P^0_{s,a}(V - \mu) - \sigma\left(\mathrm{sp}((V-\mu))_*\right) \right\} \\ &\quad - \max_{\mu^{\lambda}_{\hat{P}^0_{s,a}} \in \mathcal{M}^{\lambda}_{\hat{P}^0_{s,a}}} \left\{ \widehat{P}^0_{s,a}(V - \mu^{\lambda}_{\hat{P}^0_{s,a}}) - \sigma\left(\mathrm{sp}((V - \mu^{\lambda}_{\hat{P}^0_{s,a}})_*)\right) \right\} \Big| &(156) \\ &\leq \sum_a \pi(a|s) \Big(2\sqrt{\frac{L}{N}} \sqrt{\mathrm{Var}_{P^0_{s,a}}(V)} + \frac{L C_S \|1\|_*}{N(1-\gamma)} \Big) &(157) \\ &= 2\sqrt{\frac{L}{N}} \sqrt{\mathrm{Var}_{P^0_{s,a}}(V)} + \frac{3 L C_S \|1\|_*}{N(1-\gamma)} &(158) \end{aligned}$$

using sa-rectangular results, which gives the result.

Combining this lemma with a matrix notation, one has with probability $1-\delta$:

$$\left| \underline{\hat{P}}^{\pi^*,V} V^{\pi^*,\sigma} - \underline{P}^{\pi^*,V} V^{\pi^*,\sigma} \right| \leq 2\sqrt{\frac{L}{N}} \sqrt{\mathrm{Var}_{P^*}(V^{\star,\sigma})} + \frac{3 L C_S \|1\|_*}{N(1-\gamma)} \tag{159}$$

$$\tag{160}$$

$\square$

9.3.4 Proof of Lemma 9

Using the same argument as in (209), it holds that for any α^* solution of (**??**) or (53)

$$\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}(V^{\star,\sigma})}=\sqrt{\frac{1}{1-\gamma}}\sqrt{\sum_{t=0}^{\infty}\gamma^t\left(\underline{\widehat{P}}^{\pi^\star,V}\right)^t\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}([V^{\star,\sigma}]_{\alpha^{**}})}. \tag{161}$$

Then we can control $\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}(V^{\star,\sigma})$. Defining $V' \coloneqq V^{\star,\sigma}-\eta 1$, $\eta\in\mathbb{R}$, we use Bellman's equation in (32)) which lead to

$$\begin{aligned}
V' = V^{\star,\sigma}-\eta 1 \le V^{\star,\sigma}-\eta 1 = r_{\pi^\star}+\gamma\underline{P}^{\pi^\star,V}V^{\star,\sigma}-\eta 1 && (162)\\
= r_{\pi^\star}+\gamma P^{\pi^\star,V}[V^{\star,\sigma}-\gamma\sigma\mathrm{sp}(V^{\star,\sigma})_* -\eta 1 && (163)\\
= r'_{\pi^\star}+\gamma\underline{\widehat{P}}^{\pi^\star,V}V'+\gamma\Big(P^{\pi^\star,V}-\underline{\widehat{P}}^{\pi^\star,V}\Big)V^{\star,\sigma}-\gamma\sigma\mathrm{sp}([V^{\star,\sigma})_* && (164)\\
= r'_{\pi^\star}+\gamma\underline{\widehat{P}}^{\pi^\star,V}V'+\gamma\Big(\underline{P}^{\pi^\star,V}-\underline{\widehat{P}}^{\pi^\star,V}\Big)V^{\star,\sigma} && (165)\\
\le r'_{\pi^\star}+\gamma\underline{\widehat{P}}^{\pi^\star,V}V'+\gamma\Big(\underline{P}^{\pi^\star,V}-\underline{\widehat{P}}^{\pi^\star,V}\Big)V^{\star,\sigma} && (166)
\end{aligned}$$

where in the second line we use Lemma 3. and we define $r'_{\pi^\star}=r_{\pi^\star}-(1-\gamma)\eta<r_{\pi^\star}<1$. We obtain the same result in s-rectangular case using lemma 4 instead. Then

$$\begin{aligned}
\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}([V^{\star,\sigma}) &\overset{\text{(a)}}{=} \mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}(V') = \underline{\widehat{P}}^{\pi^\star,V}(V'\circ V') - \big(\underline{\widehat{P}}^{\pi^\star,V}V'\big)\circ\big(\underline{\widehat{P}}^{\pi^\star,V}V'\big)\\
&= \underline{\widehat{P}}^{\pi^\star,V}(V'\circ V') - \big(\underline{\widehat{P}}^{\pi^\star,V}V'\big)\circ\big(\underline{\widehat{P}}^{\pi^\star,V}V'\big)\\
&\overset{\text{(b)}}{\le} \underline{\widehat{P}}^{\pi^\star,V}(V'\circ V') - \frac{1}{\gamma^2}\Big(V'-r'_{\pi^\star}-\gamma\Big(\underline{P}^{\pi^\star,V}-\underline{\widehat{P}}^{\pi^\star,V}\Big)V^{\star,\sigma}\Big)^{\circ 2}\\
&= \underline{\widehat{P}}^{\pi^\star,V}(V'\circ V') - \frac{1}{\gamma^2}V'\circ V' + \frac{2}{\gamma^2}V'\circ\Big(r'_{\pi^\star}+\gamma\Big(\underline{P}^{\pi^\star,V}-\underline{\widehat{P}}^{\pi^\star,V}\Big)V^{\star,\sigma}\Big)\\
&\quad - \frac{1}{\gamma^2}\Big(r'_{\pi^\star}+\gamma\Big(\underline{P}^{\pi^\star,V}-\underline{\widehat{P}}^{\pi^\star,V}\Big)V^{\star,\sigma}\Big)^{\circ 2}\\
&\overset{\text{(c)}}{\le} \underline{\widehat{P}}^{\pi^\star,V}(V'\circ V') - \frac{1}{\gamma}V'\circ V' + \frac{2}{\gamma^2}\|V'\|_\infty 1 && (167)\\
&+ \frac{2}{\gamma}\|V'\|_\infty\Big|\Big(\underline{P}^{\pi^\star,V}-\underline{\widehat{P}}^{\pi^\star,V}\Big)V^{\star,\sigma}\Big| && (168)\\
&\le \underline{\widehat{P}}^{\pi^\star,V}(V'\circ V') - \frac{1}{\gamma}V'\circ V' + \frac{2}{\gamma^2}\|V'\|_\infty 1 && (169)\\
&+ \frac{2}{\gamma}\|V'\|_\infty\Big(2\sqrt{\frac{L}{(1-\gamma)^2N}}+\frac{3C_S\left\|1\right\|_* L}{N(1-\gamma)}\Big)1, && (170)
\end{aligned}$$

where (a) holds by the fact that $\mathrm{Var}_{P_\pi}(V-c1)=\mathrm{Var}_{P_\pi}(V)$ for any scalar c and $V\in\mathbb{R}^S$, (b) follows from (166), (c) arises from $\frac{1}{\gamma^2}V'\circ V'\ge\frac{1}{\gamma}V'\circ V'$ and $-1\le r_{\pi^\star}-(1-\gamma)V_{\min}1=r'_{\pi^\star}\le r_{\pi^\star}\le 1$, and the last inequality holds by Lemma 8. Plugging (170) into (161) leads to

$$
\begin{aligned}
&\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}(V^{\star,\sigma})} && (171)\\
&\leq \sqrt{\frac{1}{1-\gamma}}\Big(\sum_{t=0}^{\infty}\gamma^t\left(\underline{\widehat{P}}^{\pi^\star,V}\right)^t\left(\underline{\widehat{P}}^{\pi^\star,V}(V'\circ V')-\frac{1}{\gamma}V'\circ V'+\frac{2}{\gamma^2}\|V'\|_\infty 1\right. && (172)\\
&\left.+\frac{2}{\gamma}\|V'\|_\infty\Big(2\sqrt{\frac{L}{(1-\gamma)^2N}}+\frac{3C_S\|1\|_* L}{N(1-\gamma)}\Big)1\right)\Big)^{1/2}\\
&\overset{\text{(i)}}{\leq}\sqrt{\frac{1}{1-\gamma}}\sqrt{\left|\sum_{t=0}^{\infty}\gamma^t\left(\underline{\widehat{P}}^{\pi^\star,V}\right)^t\left(\underline{\widehat{P}}^{\pi^\star,V}(V'\circ V')-\frac{1}{\gamma}V'\circ V'\right)\right|}\\
&\quad+\sqrt{\frac{1}{1-\gamma}}\sqrt{\sum_{t=0}^{\infty}\gamma^t\left(\underline{\widehat{P}}^{\pi^\star,V}\right)^t\left(\frac{2}{\gamma^2}\|V'\|_\infty 1+\frac{2}{\gamma}\|V'\|_\infty\Big(2\sqrt{\frac{L}{(1-\gamma)^2N}}+\frac{3C_S\|1\|_* L}{N(1-\gamma)}\Big)1\right)}\\
&\leq \sqrt{\frac{1}{1-\gamma}}\sqrt{\left|\sum_{t=0}^{\infty}\gamma^t\left(\underline{\widehat{P}}^{\pi^\star,V}\right)^t\left[\underline{\widehat{P}}^{\pi^\star,V}(V'\circ V')-\frac{1}{\gamma}V'\circ V'\right]\right|} && (173)\\
&+\sqrt{\frac{\Big(2+2\Big(2\sqrt{\frac{L}{(1-\gamma)^2N}}+\frac{3C_S\|1\|_* L}{N(1-\gamma)}\Big)\Big)\|V'\|_\infty}{(1-\gamma)^2\gamma^2}}1, && (174)
\end{aligned}
$$

where (i) holds by the triangle inequality. Therefore, the remainder of the proof shall focus on the first term, which follows

$$
\begin{aligned}
&\left|\sum_{t=0}^{\infty}\gamma^t\Big(\underline{\widehat{P}}^{\pi^\star,V}\Big)^t\Big(\underline{\widehat{P}}^{\pi^\star,V}(V'\circ V')-\frac{1}{\gamma}V'\circ V'\Big)\right|\\
&=\left|\left(\sum_{t=0}^{\infty}\gamma^t\Big(\underline{\widehat{P}}^{\pi^\star,V}\Big)^{t+1}-\sum_{t=0}^{\infty}\gamma^{t-1}\Big(\underline{\widehat{P}}^{\pi^\star,V}\Big)^t\right)(V'\circ V')\right|\leq\frac{1}{\gamma}\|V'\|_\infty^2 1 \qquad (175)
\end{aligned}
$$

by recursion. Inserting (175) back to (174) leads to

$$
\begin{aligned}
&\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}(V^{\star,\sigma}]_{\alpha^*})}\\
&\leq\sqrt{\frac{\|V\|_\infty^2}{\gamma(1-\gamma)}}1+3\sqrt{\frac{\Big(1+\Big(\sqrt{\frac{L}{(1-\gamma)^2N}}+\frac{C_S\|1\|_* L}{N(1-\gamma)}\Big)\Big)\|V'\|_\infty}{(1-\gamma)^2\gamma^2}}1\\
&\leq 4\sqrt{\frac{\Big(1+\Big(\sqrt{\frac{L}{(1-\gamma)^2N}}+\frac{C_S\|1\|_* L}{N(1-\gamma)}\Big)\Big)\|V'\|_\infty}{(1-\gamma)^2\gamma^2}}1 && (176)\\
&\leq 4\sqrt{\frac{\Big(1+\Big(1\sqrt{\frac{L}{(1-\gamma)^2N}}+\frac{C_S\|1\|_* L}{N(1-\gamma)}\Big)\Big)\|V'\|_*}{(1-\gamma)^2\gamma^2}}1 && (177)
\end{aligned}
$$

Taking the infimum over η in the right-hand side, recall $V' := V^{\star,\sigma} - \eta 1$, we obtain the definition of the span semi norm.

$$\left(I - \gamma \underline{\widehat{P}}^{\pi^\star, V}\right)^{-1} \sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star, V}}([V^{\star,\sigma}]_{\alpha^*})} \leq 4\sqrt{\frac{\left(1 + \left(\sqrt{\frac{L}{(1-\gamma)^2 N}} + \frac{C_S \|1\|_* L}{N(1-\gamma)}\right)\right)\mathrm{sp}(V^{\star,\sigma})_*}{(1-\gamma)^2\gamma^2}}1$$

$$\leq 4\sqrt{\frac{\left(1 + \left(\sqrt{\frac{L}{(1-\gamma)^2 N}} + \frac{C_S \|1\|_* L}{N(1-\gamma)}\right)\right)}{\gamma^3(1-\gamma)^2 \max\{1-\gamma, C_g\sigma\}}}1 \tag{178}$$

$$\leq 4\sqrt{\frac{\left(1 + \left(\sqrt{\frac{L}{(1-\gamma)^2 N}} + \frac{C_S \|1\|_* L}{N(1-\gamma)}\right)\right)}{\gamma^3(1-\gamma)^3}}1, \tag{179}$$

where the penultimate inequality follows from applying Lemma 5 with $P = P^0$ and $\pi = \pi^\star$:

$$\mathrm{sp}(V^{\star,\sigma})_* \leq \frac{1}{\gamma \max\{1-\gamma, C_g\sigma\}}.$$

or with an extra factor for s rectangular assumptions.

$$\mathrm{sp}(V^{\star,\sigma})_* \leq \frac{1}{\gamma \max\{1-\gamma, \min_s \|\pi_s\|_* \tilde{\sigma} C g\}}.$$

9.3.5 Proof of Lemma 10

In this proof, we will sa-rectangular notations, especially $\alpha^{**}_{s,a}$ but it holds also for α^{**}_s and s-rectangular case. For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, using the results in (141), for both sa-rectangular case:

$$\left|\widehat{P}^{\widehat{\pi},\widehat{V}}_{s,a} \widehat{V}^{\widehat{\pi},\sigma} - P^{\widehat{\pi},\widehat{V}}_{s,a} \widehat{V}^{\widehat{\pi},\sigma}\right| \leq \max\left\{ \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_{\alpha^{\lambda,\omega*}_{P_{s,a}}}\right|, \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_{\alpha^{\lambda,\omega*}_{\hat{P}_{s,a}}}\right| \right\} \tag{180}$$

The first term in this max can be bounded using:

$$\left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_{\alpha^{\lambda,\omega*}_{P_{sa}}}\right| \tag{181}$$

$$\overset{(a)}{\leq} \left(\left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\star,\sigma}\right]_{\alpha^{\lambda,\omega*}_{P_{sa}}}\right| + \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_{\alpha^{\lambda,\omega*}_{P_{sa}}} - \left[\widehat{V}^{\star,\sigma}\right]_{\alpha^{\lambda,\omega*}_{P_{sa}}}\right)\right|\right)$$

$$\leq \left(\left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\star,\sigma}\right]_{\alpha^{\lambda,\omega*}_{P_{sa}}}\right| + \left\|P^0_{s,a} - \widehat{P}^0_{s,a}\right\|_1 \left\|\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_{\alpha^{\lambda,\omega*}_{P_{sa}}} - \left[\widehat{V}^{\star,\sigma}\right]_{\alpha^{\lambda,\omega*}_{P_{sa}}}\right\|_\infty\right)$$

$$\overset{(b)}{\leq} \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\star,\sigma}\right]_{\alpha^{\lambda,\omega*}_{P_{sa}}}\right| + 2\left\|\widehat{V}^{\widehat{\pi},\sigma} - \widehat{V}^{\star,\sigma}\right\|_\infty$$

$$\overset{(c)}{\leq} \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\star,\sigma}\right]_{\alpha^{\lambda,\omega*}_{P_{sa}}}\right| + 2\varepsilon_{\mathsf{opt}} \tag{182}$$

where (a) comes from the triangle inequality, and (b) comes from $\left\|P^0_{s,a} - \widehat{P}^0_{s,a}\right\|_1 \leq 2$ and $\left\|\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_{\alpha^{\lambda,\omega*}_{P_{sa}}} - \left[\widehat{V}^{\star,\sigma}\right]_{\alpha^{\lambda,\omega*}_{P_{sa}}}\right\|_\infty \leq \left\|\widehat{V}^{\widehat{\pi},\sigma} - \widehat{V}^{\star,\sigma}\right\|_\infty$, and (c) follows from the definition of the optimization error in (55). The second term of the max can be controlled in the same manner, i.e.:

$$\left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_{\alpha^{\lambda,\omega*}_{\hat{P}_{sa}}}\right| \leq \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\star,\sigma}\right]_{\alpha^{\lambda,\omega*}_{\hat{P}_{sa}}}\right| + 2\varepsilon_{\mathsf{opt}} \tag{183}$$

$$\leq |\max_{\mu^\lambda_{P^0_{s,a}} \in \mathcal{M}^\lambda_{P^0_{s,a}}} \left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)(\widehat{V}^{\star,\sigma} - \mu^\lambda_{P^0_{s,a}}) + \max_{\mu^\lambda_{\hat{P}^0_{s,a}} \in \mathcal{M}^\lambda_{\hat{P}^0_{s,a}}} \left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)(\mu^\lambda_{P^0_{s,a}} - \mu^\lambda_{\hat{P}^0_{s,a}})| \tag{184}$$

$$+ 2\varepsilon_{\mathsf{opt}} \tag{185}$$

where the last inequality follow the decomposition of (147). Finally, to control the remaining term

$$\max_{\mu^{\lambda}_{P^0_{s,a}} \in \mathcal{M}^{\lambda}_{P^0_{s,a}}} \left(P^0_{s,a} - \widehat{P}^0_{s,a} \right) (\widehat{V}^{\star,\sigma} - \mu^{\lambda}_{P^0_{s,a}}) = \max_{\alpha^{\lambda}_P \in \mathrm{A}^{\lambda}_P} \left\{ (P^0_{s,a} - \widehat{P}^0_{s,a}) \, [V]_{\alpha^{\lambda}_P} \right\} \tag{186}$$

(185) for any given $\alpha \in [0, \alpha^{\lambda,\omega*}_{P_{sa}}[\subset \left[0, \frac{1}{1-\gamma}\right]^S$ in the variational family with one parameter λ, with the dependency between $\widehat{V}^{\star,\sigma}$ and $\widehat{P}^0$, we resort to the following leave-one-out argument or absorbing MDPs used in [Agarwal et al., 2020, Li et al., 2022b, Shi and Chi, 2022, Clavier et al., 2023]. To begin, we create a collection of auxiliary RMDPs that exhibit the intended statistical independence between robust value functions and the estimated nominal transition kernel. These auxiliary RMDPs are designed to be minimally distinct from the initial RMDPs, subsequently, we manage to control the relevant term within these auxiliary RMDPs and demonstrate that its value closely approximates the target quantity for the desired RMDP. Recall that the empirical infinite-horizon robust MDP $\widehat{\mathcal{M}}_{\mathsf{rob}}$ is defined using the nominal transition kernel $\widehat{P}^0$. Inspired by Agarwal et al. [2020], we can construct an auxiliary absorbing robust MDP $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ for each state s and any non-negative scalar $u \geq 0$, so that it is the same as $\widehat{\mathcal{M}}_{\mathsf{rob}}$ except for the transition properties in state s. These auxiliary MDPS are called absorbing MDPs are have been used for the first time in the context of RMDPS in Clavier et al. [2023]. Defining the reward function and nominal transition kernel of $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ as $P^{s,u}$ and $r^{s,u}$, which are expressed as follows using the same notation as Shi et al. [2023]:

$$\begin{cases} r^{s,u}(s,a) = u & \forall a \in \mathcal{A}, \\ r^{s,u}(\widetilde{s},a) = r(\widetilde{s},a) & \forall (\widetilde{s},a) \in \mathcal{S} \times \mathcal{A} \text{ and } \widetilde{s} \neq s. \end{cases} \tag{187}$$

$$\begin{cases} P^{s,u}(s' \,|\, s,a) = \mathbb{1}(s' = s) & \forall (s',a) \in \mathcal{S} \times \mathcal{A}, \\ P^{s,u}(\cdot \,|\, \widetilde{s},a) = \widehat{P}^0(\cdot \,|\, \widetilde{s},a) & \forall (\widetilde{s},a) \in \mathcal{S} \times \mathcal{A} \text{ and } \widetilde{s} \neq s, \end{cases} \tag{188}$$

Nominal transition probability at state s of the auxiliary $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ never leaves state s once entered, which gives the name absorbing to these auxiliary RMPDs. Finally, we define the robust Bellman operator $\widehat{\mathcal{T}}^{\sigma}_{s,u}(\cdot)$ associated $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ as

$$\widehat{\mathcal{T}}^{\sigma}_{s,u}(Q)(\tilde{s},a) = r^{s,u}(\tilde{s},a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{sa},\sigma}(P^{s,u}_{\tilde{s},a})} \mathcal{P}V, \qquad \text{with } V(\tilde{s}) = \max_a Q(\tilde{s},a). \tag{189}$$

in sa-rectangular case and with stochastic policy in s-rectangular case. Using these auxiliary RMDPs we can remark equivalence between $\widehat{\mathcal{M}}_{\mathsf{rob}}$ and the auxiliary RMDP $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ fixed-point. First, $\widehat{Q}^{\star,\sigma}$ is the unique-fixed point of $\widehat{\mathcal{T}}^{\sigma}(\cdot)$ with associated value $\widehat{V}^{\star,\sigma}$. We will show that the robust value function $\widehat{V}^{\star,\sigma}_{s,u^\star}$ obtained from the fixed point of $\widehat{\mathcal{T}}^{\sigma}_{s,u}(\cdot)$is the same as the the robust value function $\widehat{V}^{\star,\sigma}$ derived from $\widehat{\mathcal{T}}^{\sigma}(\cdot)$, as long as we choose u as

$$u^\star := u^\star(s) = \widehat{V}^{\star,\sigma}(s) - \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{sa},\sigma}(e_s)} \mathcal{P}\widehat{V}^{\star,\sigma}. \tag{190}$$

with e_s is the s-th standard basis vector in $\mathbb{R}^S$. This assertion is verified as:

- **First for state $s' \neq s$, for all $a \in \mathcal{A}$**: it holds

$$\begin{aligned} r^{s,u^\star}(s',a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{sa},\sigma}(P^{s,u^\star}_{s',a})} \mathcal{P}\widehat{V}^{\star,\sigma} &= r(s',a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{sa},\sigma}(\widehat{P}^0_{s',a})} \mathcal{P}\widehat{V}^{\star,\sigma} \\ &= \widehat{\mathcal{T}}^{\sigma}(\widehat{Q}^{\star,\sigma})(s',a) = \widehat{Q}^{\star,\sigma}(s',a), \end{aligned} \tag{191}$$

  where the first equality holds because of (187) and (188), and the last inequality comes from that $\widehat{Q}^{\star,\sigma}$ is the fixed point of $\widehat{\mathcal{T}}^{\sigma}(\cdot)$ (see Lemma 8.3) and the definition of the robust Bellman operator in (13).
- **Then for state s, for any $a \in \mathcal{A}$** :

$$\begin{aligned} r^{s,u^\star}(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(P^{s,u^\star}_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma} &= u^\star + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{sa},\sigma}(e_s)} \mathcal{P}\widehat{V}^{\star,\sigma} \\ &= \widehat{V}^{\star,\sigma}(s) - \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{sa},\sigma}(e_s)} \mathcal{P}\widehat{V}^{\star,\sigma} + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{sa},\sigma}(e_s)} \mathcal{P}\widehat{V}^{\star,\sigma} = \widehat{V}^{\star,\sigma}(s), \end{aligned} \tag{192}$$

using in the first equality is the definition of $P_{s,a}^{s,u^\star}$ in (188) and where we use the definition of $u^\star$ in (190) in the second one.

Finally, we have proved that there exists a fixed point $\widehat{Q}_{s,u^\star}^{\star,\sigma}$ of the operator $\widehat{\mathcal{T}}_{s,u^\star}^{\sigma}(\cdot)$ by taking

$$\begin{cases} \widehat{Q}_{s,u^\star}^{\star,\sigma}(s,a) = \widehat{V}^{\star,\sigma}(s) & \forall a \in \mathcal{A}, \\ \widehat{Q}_{s,u^\star}^{\star,\sigma}(s',a) = \widehat{Q}^{\star,\sigma}(s',a) & \forall s' \neq s \text{ and } a \in \mathcal{A}. \end{cases} \tag{193}$$

we have confirmed the existence of a fixed point of the operator $\widehat{\mathcal{T}}_{s,u^\star}^{\sigma}(\cdot)$ with corresponding value function $\widehat{V}_{s,u^\star}^{\star,\sigma}$ that coincide with $\widehat{V}^{\star,\sigma}$. Note that the corresponding properties between $\widehat{\mathcal{M}}_{\mathsf{rob}}$ and $\widehat{\mathcal{M}}_{\mathsf{rob}}^{s,u}$ in Step 1 and Step 2 hold in fact for any uncertainty set and s- or sa-rectangular assumptions. Equipped with these fixed point equalities, we can use concentration inequalities to show this lemma.

Concentration inequality using an ε-net for all reward values u. First we can verify that

$$0 \leq u^\star \leq [\widehat{V}^{\star,\sigma}(s)]_{\alpha_{P_{s,a}}^{\lambda,\omega*}} \leq \widehat{V}^{\star,\sigma}(s) \leq \frac{1}{1-\gamma}. \tag{194}$$

We first construct a N_{ε_2}-net over the interval $\left[0, 1/(1-\gamma)\right]$, where $|N_{\varepsilon_2}|$ the size of the net can be controlled by $|N_{\varepsilon_2}| \leq \frac{3}{\varepsilon_2(1-\gamma)}$ [Vershynin, 2018]. The only parameter that vary is λ in the variation family $\alpha_{P_{sa}}^{\lambda}$ so we have 1-dimensional control and not a vector in $\mathbb{R}^S$. Then similarly to Lemma 8.3, it holds that for each $u \in N_{\varepsilon_2}$, there exists a unique fixed point $\widehat{Q}_{s,u}^{\star,\sigma}$ of the operator $\widehat{\mathcal{T}}_{s,u}^{\sigma}(\cdot)$, which satisfies $0 \leq \widehat{Q}_{s,u}^{\star,\sigma} \leq \frac{1}{1-\gamma} \cdot 1$. Consequently, the corresponding robust value function can be upper bounded by $\left\|\widehat{V}_{s,u}^{\star,\sigma}\right\|_\infty \leq \frac{1}{1-\gamma}$. Using (188) and (187) by construction for all $u \in N_{\varepsilon_2}$, $\widehat{\mathcal{M}}_{\mathsf{rob}}^{s,u}$ is statistically independent of $\widehat{P}_{s,a}^0$. This independence indicates that $[\widehat{V}_{s,u}^{\star,\sigma}]_\alpha$ and $\widehat{P}_{s,a}^0$ are independent for a fixed α. Using (145) and (146) and taking the union bound over all $(s,a,\alpha) \in \mathcal{S} \times \mathcal{A} \times N_{\varepsilon_1}$, $u \in N_{\varepsilon_2}$ gives that, with probability at least $1-\delta$, it holds for all $(s,a,u) \in \mathcal{S} \times \mathcal{A} \times N_{\varepsilon_2}$ that

$$\begin{aligned} \max_{\alpha_{P_{sa}}^{\lambda,\omega} \in \mathrm{A}_{P_{sa}}^{\lambda,\omega}} \left| \left(P_{s,a}^0 - \widehat{P}_{s,a}^0\right) \left[\widehat{V}_{s,u}^{\star,\sigma}\right]_{\alpha_{P_{sa}}^{\lambda,\omega*}} \right| &\leq 2\sqrt{\frac{2\log(\frac{18\|1\|_* SAN|N_{\varepsilon_2}|}{\delta})}{N}} \sqrt{\mathrm{Var}_{P_{s,a}^0}(\widehat{V}_{s,u}^{\star,\sigma})} \qquad (195) \\ &\quad + \varepsilon_2 \\ &\leq 2\sqrt{\frac{2\log(\frac{18\|1\|_* SAN|N_{\varepsilon_2}|}{\delta})}{(1-\gamma)^2 N}} + \varepsilon_2, \qquad (196) \end{aligned}$$

Finally, we use **uniform concentration** to obtain the lemma. Recalling that $u^\star \in \left[0, \frac{1}{1-\gamma}\right]$ (see (194)), we can always find some $\overline{u} \in N_{\varepsilon_2}$ such that $|\overline{u} - u^\star| \leq \varepsilon_2$. Consequently, plugging in the operator $\widehat{\mathcal{T}}_{s,u}^{\sigma}(\cdot)$ in (189) yields

$$\forall Q \in \mathbb{R}^{SA}: \quad \left\|\widehat{\mathcal{T}}_{s,\overline{u}}^{\sigma}(Q) - \widehat{\mathcal{T}}_{s,u^\star}^{\sigma}(Q)\right\|_\infty = |\overline{u} - u^\star| \leq \varepsilon_2$$

We can then remark that the fixed points of $\widehat{\mathcal{T}}_{s,\overline{u}}^{\sigma}(\cdot)$ and $\widehat{\mathcal{T}}_{s,u^\star}^{\sigma}(\cdot)$ obey

$$\begin{aligned} \left\|\widehat{Q}_{s,\overline{u}}^{\star,\sigma} - \widehat{Q}_{s,u^\star}^{\star,\sigma}\right\|_\infty &= \left\|\widehat{\mathcal{T}}_{s,\overline{u}}^{\sigma}(\widehat{Q}_{s,\overline{u}}^{\star,\sigma}) - \widehat{\mathcal{T}}_{s,u^\star}^{\sigma}(\widehat{Q}_{s,u^\star}^{\star,\sigma})\right\|_\infty \\ &\leq \left\|\widehat{\mathcal{T}}_{s,\overline{u}}^{\sigma}(\widehat{Q}_{s,\overline{u}}^{\star,\sigma}) - \widehat{\mathcal{T}}_{s,\overline{u}}^{\sigma}(\widehat{Q}_{s,u^\star}^{\star,\sigma})\right\|_\infty + \left\|\widehat{\mathcal{T}}_{s,\overline{u}}^{\sigma}(\widehat{Q}_{s,u^\star}^{\star,\sigma}) - \widehat{\mathcal{T}}_{s,u^\star}^{\sigma}(\widehat{Q}_{s,u^\star}^{\star,\sigma})\right\|_\infty \\ &\leq \gamma \left\|\widehat{Q}_{s,\overline{u}}^{\star,\sigma} - \widehat{Q}_{s,u^\star}^{\star,\sigma}\right\|_\infty + \varepsilon_2, \end{aligned}$$

where we use that the operator $\widehat{\mathcal{T}}_{s,u}^{\sigma}(\cdot)$ is a γ-contraction. It gives that:

$$\left\|\widehat{Q}_{s,\overline{u}}^{\star,\sigma} - \widehat{Q}_{s,u^\star}^{\star,\sigma}\right\|_\infty \leq \frac{\varepsilon_2}{(1-\gamma)} \quad \text{and} \quad \left\|\widehat{V}_{s,\overline{u}}^{\star,\sigma} - \widehat{V}_{s,u^\star}^{\star,\sigma}\right\|_\infty \leq \left\|\widehat{Q}_{s,\overline{u}}^{\star,\sigma} - \widehat{Q}_{s,u^\star}^{\star,\sigma}\right\|_\infty \leq \frac{\varepsilon_2}{(1-\gamma)}. \tag{197}$$

Finally to control the first term in (185), using the identity $\widehat{V}^{\star,\sigma} = \widehat{V}^{\star,\sigma}_{s,u^\star}$ or fixed point relation between the two RMPDS, established in previous step of the proof gives that: for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$
\begin{aligned}
&\max_{\alpha^{\lambda,\omega}_{P_{s,a}} \in \mathrm{A}^{\lambda,\omega}_{P_{s,a}}} \left| \left(P^0_{s,a} - \widehat{P}^0_{s,a}\right) [\widehat{V}^{\star,\sigma}]_{\alpha^{\lambda,\omega}_{P_{s,a}}} \right| \\
&\leq \max_{\alpha^{\lambda,\omega}_{P_{s,a}} \in \mathrm{A}^{\lambda,\omega}_{P_{s,a}}} \left| \left(P^0_{s,a} - \widehat{P}^0_{s,a}\right) [\widehat{V}^{\star,\sigma}]_{\alpha^{\lambda,\omega}_{P_{s,a}}} \right| \\
&\overset{(a)}{\leq} \max_{\alpha^{\lambda,\omega}_{P_{s,a}} \in \mathrm{A}^{\lambda,\omega}_{P_{s,a}}} \left\{ \left| \left(P^0_{s,a} - \widehat{P}^0_{s,a}\right) [\widehat{V}^{\star,\sigma}_{s,\overline{u}}]_{\alpha^{\lambda,\omega}_{P_{s,a}}} \right| + \left| \left(P^0_{s,a} - \widehat{P}^0_{s,a}\right) \left([\widehat{V}^{\star,\sigma}_{s,\overline{u}}]_{\alpha^{\lambda,\omega}_{P_{s,a}}} - [\widehat{V}^{\star,\sigma}_{s,u^\star}]_{\alpha^{\lambda}_{P_{s,a}}} \right) \right| \right\} \\
&\overset{(b)}{\leq} \max_{\alpha^{\lambda,\omega}_{P_{s,a}} \in \mathrm{A}^{\lambda,\omega}_{P_{s,a}}} \left| \left(P^0_{s,a} - \widehat{P}^0_{s,a}\right) [\widehat{V}^{\star,\sigma}_{s,\overline{u}}]_{\alpha^{\lambda}_{P_{s,a}}} \right| + \frac{2\varepsilon_2}{(1-\gamma)} \\
&\overset{(c)}{\leq} \frac{2\varepsilon_2}{(1-\gamma)} + \varepsilon_2 + 2\sqrt{\frac{2\log(\frac{18\|1\|_* SAN|N_{\varepsilon_2}|}{\delta})}{N}} \sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma}_{s,\overline{u}})} + \frac{4\log(\frac{18\|1\|_* SAN|N_{\varepsilon_2}|}{\delta})}{3N(1-\gamma)} \\
&\leq \frac{3\varepsilon_2}{(1-\gamma)} + 2\sqrt{\frac{2\log(\frac{18\|1\|_* SAN|N_{\varepsilon_2}|}{\delta})}{N}} \sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma})} + \frac{4\log(\frac{18\|1\|_* SAN|N_{\varepsilon_2}|}{\delta})}{3N(1-\gamma)} \\
&\quad + 2\sqrt{\frac{2\log(\frac{18\|1\|_* SAN|N_{\varepsilon_2}|}{\delta})}{N}} \sqrt{\left|\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma}) - \mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma}_{s,\overline{u}})\right|} \\
&\overset{(d)}{\leq} \frac{3\varepsilon_2}{(1-\gamma)} + 2\sqrt{2\frac{\log(\frac{18\|1\|_* SAN|N_{\varepsilon_2}|}{\delta})}{N}} \sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma})} + 2\sqrt{\frac{4\varepsilon_2 \log(\frac{18\|1\|_* SAN|N_{\varepsilon_2}|}{\delta})}{N(1-\gamma)^2}} && (198) \\
&\leq 2\sqrt{\frac{L''}{N}} \sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma})} + \frac{14\log(\frac{54\|1\|_* SAN|N_{\varepsilon_2}|}{\delta})}{N(1-\gamma)} && (199) \\
&\leq 16\sqrt{\frac{L''}{(1-\gamma)^2 N}}, && (200)
\end{aligned}
$$

with $L'' = \log\left(\frac{54\|1\|_* SAN^2}{(1-\gamma)\delta}\right)$ where (a) comes from triangular inequality, (b) is due (197), for any $\alpha \in \mathbb{R}^S$

$$
\begin{aligned}
\left| \left(P^0_{s,a} - \widehat{P}^0_{s,a}\right) \left([\widehat{V}^{\star,\sigma}_{s,\overline{u}}]_\alpha - [\widehat{V}^{\star,\sigma}_{s,u^\star}]_\alpha \right) \right| &\leq \left\| P^0_{s,a} - \widehat{P}^0_{s,a} \right\|_1 \left\| [\widehat{V}^{\star,\sigma}_{s,\overline{u}}]_\alpha - [\widehat{V}^{\star,\sigma}_{s,u^\star}]_\alpha \right\|_\infty \\
&\leq 2\left\| \widehat{V}^{\star,\sigma}_{s,\overline{u}} - \widehat{V}^{\star,\sigma}_{s,u^\star} \right\|_\infty \leq \frac{2\varepsilon_2}{(1-\gamma)}, \qquad (201)
\end{aligned}
$$

(c) follows from (195), (d) holds using Lemma 1 with (197). Here, the two last inequalities hold by letting $\varepsilon_2 = \frac{2\log(\frac{18\|1\|_* SAN|N_{\varepsilon_2}|}{\delta})}{N}$, which gives $|N_{\varepsilon_2}| \leq \frac{3}{\varepsilon_2(1-\gamma)} \leq \frac{3N}{1-\gamma}$, and the last inequality holds by the fact $\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma}) \leq \|\widehat{V}^{\star,\sigma}\|_\infty \leq \frac{1}{1-\gamma}$ and letting $N \geq 2\log\left(\frac{54\|1\|_* SAN^2}{(1-\gamma)\delta}\right) = L''$.

Rewriting (180), the first term of the max is controlled.

$$
\max\left\{ \left| \left(P^0_{s,a} - \widehat{P}^0_{s,a}\right) [\widehat{V}^{\widehat{\pi},\sigma}]_{\alpha^{\lambda^*}_{P_{s,a}}} \right|, \left| \left(P^0_{s,a} - \widehat{P}^0_{s,a}\right) [\widehat{V}^{\widehat{\pi},\sigma}]_{\alpha^{\lambda^*}_{\hat{P}_{s,a}}} \right| \right\}
$$

The second term can be controlled by the same term as the first one plus an additional term with

$$
\begin{aligned}
&\left| \left(P^0_{s,a} - \widehat{P}^0_{s,a}\right) [\widehat{V}^{\widehat{\pi},\sigma}]_{\alpha^{\lambda^*}_{\hat{P}_{s,a}}} \right| \leq \\
&|\max_{\mu^{\lambda}_{P^0_{s,a}} \in \mathcal{M}^{\lambda}_{P^0_{s,a}}} \left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)(\widehat{V}^{\star,\sigma} - \mu^{\lambda}_{P^0_{s,a}}) + \max_{\mu^{\lambda}_{\hat{P}^0_{s,a}} \in \mathcal{M}^{\lambda}_{\hat{P}^0_{s,a}}} \left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)(\mu^{\lambda}_{P^0_{s,a}} - \mu^{\lambda}_{\hat{P}^0_{s,a}})|
\end{aligned}
$$

and similarly to previous lemma in (153), the residual or term in the right in the previous equation can be controlled with $\frac{L' C_S \|1\|_*}{N(1-\gamma)}$ Finally, putting (199) and (200) back into Equation (185) and using Eq. (200) with probability at least $1-\delta$ we obtain

$$\begin{aligned}\left|\widehat{P}^{\widehat{\pi},\widehat{V}}_{s,a}\widehat{V}^{\widehat{\pi},\sigma} - P^{\widehat{\pi},\widehat{V}}_{s,a}\widehat{V}^{\widehat{\pi},\sigma}\right| &\leq \max_{\alpha^{\lambda,\omega}_{P_{s,a}} \in \mathrm{A}^{\lambda,\omega}_{P_{s,a}}} \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[\widehat{V}^{\star,\sigma}]_{\alpha^{\lambda,\omega}_{P_{s,a}}}\right| + 2\varepsilon_{\mathsf{opt}} \\ &\leq 2\sqrt{\frac{L'}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma})} + 2\varepsilon_{\mathsf{opt}} + \frac{14L'' C_S \|1\|_*}{N(1-\gamma)} \\ &\leq 2\sqrt{\frac{L''}{(1-\gamma)^2 N}} + 2\varepsilon_{\mathsf{opt}} + \frac{14L'' C_S \|1\|_*}{N(1-\gamma)},\end{aligned} \tag{202}$$

$\forall (s,a) \in \mathcal{S}\times\mathcal{A}$. Using matrix form we obtain finally:

$$\begin{aligned}\left|\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right| &\leq 2\sqrt{\frac{L''}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma})}1 + 2\varepsilon_{\mathsf{opt}}1 \\ &\leq 2\sqrt{\frac{L''}{(1-\gamma)^2 N}}1 + 2\varepsilon_{\mathsf{opt}}1. + \frac{14L'' C_S \|1\|_*}{N(1-\gamma)}\end{aligned} \tag{203}$$

The proof is similar in the s-rectangular case, factorising by $\pi(a|s)$, like in in 8. Moreover, the proof is similar for TV without the geometric term depending on C_S.

9.3.6 Proof of Lemma 11

We always use the same manner as in Appendix 9.3.4. Similarly to (161), it holds:

$$\left(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})} \leq \sqrt{\frac{1}{1-\gamma}}\sqrt{\sum_{t=0}^{\infty}\gamma^t\left(\underline{P}^{\widehat{\pi},\widehat{V}}\right)^t \mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})}. \tag{204}$$

In order to upper bound $\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})$, we define $V' := \widehat{V}^{\widehat{\pi},\sigma} - \eta 1$ for any α^* solving a dual optimization problem with $\eta \in \mathbb{R}$. Using as (168), it holds

$$\begin{aligned}&\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma}) \leq \underline{P}^{\widehat{\pi},\widehat{V}}(V'\circ V') - \frac{1}{\gamma}V'\circ V' + \frac{2}{\gamma^2}\|V'\|_\infty 1 + \frac{2}{\gamma}\|V'\|_\infty\left|\left(\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}} - \underline{P}^{\widehat{\pi},\widehat{V}}\right)\widehat{V}^{\widehat{\pi},\sigma}\right| \\ &\leq \underline{P}^{\widehat{\pi},\widehat{V}}(V'\circ V') - \frac{1}{\gamma}V'\circ V' + \frac{2}{\gamma^2}\|V'\|_\infty 1 + \frac{2}{\gamma}\|V'\|_\infty\left(2\sqrt{\frac{L''}{(1-\gamma)^2 N}} + 2\varepsilon_{\mathsf{opt}} + \frac{14L'' C_S \|1\|_*}{N(1-\gamma)}\right)1,\end{aligned} \tag{205}$$

where the last inequality makes use of Lemma 10. Plugging (205) back into (204) leads to

$$
\begin{aligned}
\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})} &\overset{\text{(a)}}{\leq} \sqrt{\frac{1}{1-\gamma}}\sqrt{\left|\sum_{t=0}^{\infty}\gamma^t\left(\underline{P}^{\widehat{\pi},\widehat{V}}\right)^t\left(\underline{P}^{\widehat{\pi},\widehat{V}}\left(V'\circ V'\right)-\frac{1}{\gamma}V'\circ V'\right)\right|} \\
&\quad+\sqrt{\frac{1}{(1-\gamma)^2\gamma^2}\left(2\sqrt{\frac{L''}{(1-\gamma)^2N}}+2\varepsilon_{\mathsf{opt}}+\frac{14L''C_S\left\|1\right\|_*}{N(1-\gamma)}\right)\|V'\|_\infty 1} \\
&\overset{\text{(b)}}{\leq} \sqrt{\frac{\|V'\|_\infty^2}{\gamma(1-\gamma)}}1+\sqrt{\frac{\left(2\sqrt{\frac{L''}{(1-\gamma)^2N}}+2\varepsilon_{\mathsf{opt}}+\frac{14L''C_S\|1\|_*}{N(1-\gamma)}\right)\|V'\|_\infty}{(1-\gamma)^2\gamma^2}}1 \\
&\overset{\text{(c)}}{\leq} \sqrt{\frac{\|V'\|_\infty^2}{\gamma(1-\gamma)}}1+5\sqrt{\left(1+\varepsilon_{\mathsf{opt}}+\frac{L''C_S\left\|1\right\|_*}{N(1-\gamma)}\right)\frac{\|V'\|_\infty}{(1-\gamma)^2\gamma^2}1} \qquad (206) \\
&\leq 6\sqrt{\left(1+\varepsilon_{\mathsf{opt}}+\frac{L''C_S\left\|1\right\|_*}{N(1-\gamma)}\right)\frac{\|V'\|_\infty}{(1-\gamma)^2\gamma^2}1}, \qquad (207)
\end{aligned}
$$

where (a) is the same as (174), (b) holds by repeating the argument of (175), (c) follows by taking $N \geq \frac{L''}{(1-\gamma)^2}$ and then the last inequality holds by $\|V'\|_\infty \leq \|V^{\star,\sigma}\|_\infty \leq \frac{1}{1-\gamma}$. Then taking the infimum over η in the right-hand side of the equation in the definition of V' and using $\mathrm{sp}(.)_\infty \leq \|.\|_*$ gives

$$
\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})} \leq 6\sqrt{\left(1+\varepsilon_{\mathsf{opt}}+\frac{L''C_S\left\|1\right\|_*}{N(1-\gamma)}\right)\frac{\mathrm{sp}(V)_\infty}{(1-\gamma)^2\gamma^2}1}
$$

Finally, applying Lemma 5 with $P=\widehat{P}^0$ and $\pi=\widehat{\pi}$ yields

$$
\mathrm{sp}(\widehat{V}^{\widehat{\pi},\sigma})_* \leq \frac{1}{\gamma\max\{1-\gamma,\gamma C_g\sigma\}}, \qquad (208)
$$

for sa-rectangular or

$$
\mathrm{sp}(\widehat{V}^{\widehat{\pi},\sigma})_* \leq \frac{1}{\gamma\max\{1-\gamma,\min_s\left\|\hat{\pi}\right\|_*\tilde{\sigma}\}}
$$

which can be inserted into (207) and gives in sa-rectangular case:

$$
\begin{aligned}
\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})} &\leq 6\sqrt{\frac{\left(1+\varepsilon_{\mathsf{opt}}+\frac{L''C_S\|1\|_*}{N(1-\gamma)}\right)}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}}}1 \\
&\leq 6\sqrt{\frac{\left(1+\varepsilon_{\mathsf{opt}}+\frac{L''C_S\|1\|_*}{N(1-\gamma)}\right)}{(1-\gamma)^3\gamma^3}}1.
\end{aligned}
$$

where first inequalities comes from that we can bound it Eq. left-hand side of equation (207) by $\|V'\|_\infty \leq \|V^{\star,\sigma}\|_\infty \leq \frac{1}{1-\gamma}$. Proof for s-rectangular is similar, but requires adding an extra factor depending on the norm of the current policy and we have:

$$
\begin{aligned}
\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})} &\leq 6\sqrt{\frac{\left(1+\varepsilon_{\mathsf{opt}}+\frac{L''C_S\|1\|_*}{N(1-\gamma)}\right)}{\gamma^3(1-\gamma)^2\max\{1-\gamma,C_g\tilde{\sigma}\min_s\left\|\hat{\pi}_s\right\|_\infty\}}}1 \\
&\leq 6\sqrt{\frac{\left(1+\varepsilon_{\mathsf{opt}}+\frac{L''C_S\|1\|_*}{N(1-\gamma)}\right)}{(1-\gamma)^3\gamma^2}}1.
\end{aligned}
$$

9.3.7 Proof of Lemma 7

Observing that each row of P_π belongs to $\Delta(S)$, it can be directly verified that each row of $(1-\gamma)\left(I-\gamma P_\pi\right)^{-1}$ falls into $\Delta(S)$. As a result,

$$\begin{aligned}
\left(I-\gamma P_\pi\right)^{-1}\sqrt{\mathrm{Var}_{P_\pi}(V^{\pi,P})} &= \frac{1}{1-\gamma}(1-\gamma)\left(I-\gamma P_\pi\right)^{-1}\sqrt{\mathrm{Var}_{P_\pi}(V^{\pi,P})} \\
&\overset{\text{(a)}}{\leq} \frac{1}{1-\gamma}\sqrt{(1-\gamma)\left(I-\gamma P_\pi\right)^{-1}\mathrm{Var}_{P_\pi}(V^{\pi,P})} \\
&= \sqrt{\frac{1}{1-\gamma}}\sqrt{\sum_{t=0}^{\infty}\gamma^t\left(P_\pi\right)^t\mathrm{Var}_{P_\pi}(V^{\pi,P})},
\end{aligned} \tag{209}$$

where (a) is due to Jensen's inequality. Then for any $\eta \in \mathbb{R}^+$, $V' := V^{\pi,P} - \eta 1$ for any α solving a dual optimization problem, we can upper bound $\mathrm{Var}_{P_\pi}(V^{\pi,P})$:

$$\begin{aligned}
&\mathrm{Var}_{P_\pi}(V^{\pi,P}) \overset{\text{(i)}}{=} \mathrm{Var}_{P_\pi}(V') = P_\pi\left(V'\circ V'\right) - \left(P_\pi V'\right)\circ\left(P_\pi V'\right) \\
&\overset{\text{(ii)}}{\leq} P_\pi\left(V'\circ V'\right) - \frac{1}{\gamma^2}\left(V' - r_\pi + (1-\gamma)\eta 1\right)\circ\left(V' - r_\pi + (1-\gamma)\eta 1\right) \\
&= P_\pi\left(V'\circ V'\right) - \frac{1}{\gamma^2}V'\circ V' + \frac{2}{\gamma^2}V'\circ\left(r_\pi - (1-\gamma)\eta 1\right) - \frac{1}{\gamma^2}\left(r_\pi - (1-\gamma)\eta 1\right)\circ\left(r_\pi - (1-\gamma)\eta 1\right) \\
&\leq P_\pi\left(V'\circ V'\right) - \frac{1}{\gamma}V'\circ V' + \frac{2}{\gamma^2}\|V'\|_\infty 1 \leq P_\pi\left(V'\circ V'\right) - \frac{1}{\gamma}V'\circ V' + \frac{2}{\gamma^2}\|V'\|_\infty 1,
\end{aligned} \tag{210}$$

where (i) holds by the fact that $\mathrm{Var}_{P_\pi}(V^{\pi,P} - b1) = \mathrm{Var}_{P_\pi}([V^{\pi,P})$ for any scalar b and $V^{\pi,P} \in \mathbb{R}^S$, (ii) follows from $V' \leq r_\pi + \gamma P_\pi V^{\pi,P} - \eta 1 = r_\pi - (1-\gamma)\eta 1 + \gamma P_\pi V'$, and the last line arises from $\frac{1}{\gamma^2}V'\circ V' \geq \frac{1}{\gamma}V'\circ V'$ and $\|r_\pi - (1-\gamma)\eta 1\|_\infty \leq 1$. for $\eta \in [0, 1/(1-\gamma)[$ Plugging (210) back to (209) leads to

$$\begin{aligned}
&\left(I-\gamma P_\pi\right)^{-1}\sqrt{\mathrm{Var}_{P_\pi}(V^{\pi,P})} \leq \sqrt{\frac{1}{1-\gamma}}\sqrt{\sum_{t=0}^{\infty}\gamma^t\left(P_\pi\right)^t\left(P_\pi\left(V'\circ V'\right) - \frac{1}{\gamma}V'\circ V' + \frac{2}{\gamma^2}\|V'\|_\infty 1\right)} \\
&\overset{\text{(i)}}{\leq} \sqrt{\frac{1}{1-\gamma}}\sqrt{\left|\sum_{t=0}^{\infty}\gamma^t\left(P_\pi\right)^t\left(P_\pi\left(V'\circ V'\right) - \frac{1}{\gamma}V'\circ V'\right)\right|} + \sqrt{\frac{1}{1-\gamma}}\sqrt{\sum_{t=0}^{\infty}\gamma^t\left(P_\pi\right)^t\frac{2}{\gamma^2}\|V'\|_\infty 1} \\
&\leq \sqrt{\frac{1}{1-\gamma}}\sqrt{\left|\left(\sum_{t=0}^{\infty}\gamma^t\left(P_\pi\right)^{t+1} - \sum_{t=0}^{\infty}\gamma^{t-1}\left(P_\pi\right)^t\right)\left(V'\circ V'\right)\right|} + \sqrt{\frac{2\|V'\|_\infty 1}{\gamma^2(1-\gamma)^2}} \\
&\overset{\text{(ii)}}{\leq} \sqrt{\frac{\|V'\|_\infty^2 1}{\gamma(1-\gamma)}} + \sqrt{\frac{2\|V'\|_\infty 1}{\gamma^2(1-\gamma)^2}} \\
&\leq \sqrt{\frac{8\|V'\|_\infty 1}{\gamma^2(1-\gamma)^2}},
\end{aligned} \tag{211}$$

$$\tag{212}$$

where (i) holds by the triangle inequality, (ii) holds by following recursion, and the last inequality holds by $\|V'\|_\infty \leq \frac{1}{1-\gamma}$. Then taking the minimum over η in the right-hand side of the equation gives the result.

$$\left(I-\gamma P_\pi\right)^{-1}\sqrt{\mathrm{Var}_{P_\pi}(V^{\pi,P})} \leq \sqrt{\frac{8\mathrm{sp}(V^{\pi,P})_\infty}{\gamma^2(1-\gamma)^2}}$$

However, we also $\|V'\|_\infty \leq \|V^{\pi,P}\|_\infty \leq \frac{1}{1-\gamma}$ in (211). So finally, the result is

$$\left(I-\gamma P_\pi\right)^{-1}\sqrt{\mathrm{Var}_{P_\pi}(V^{\pi,P})} \leq \sqrt{\frac{8}{\gamma^2(1-\gamma)^2}\min\{\mathrm{sp}([V^{\pi,P})_\infty, \frac{1}{1-\gamma}\}1}.$$

10 Proof of Theorem 2

In this section, we focus on the scenarios in the uncertainty sets are constructed with (s,a)-rectangularity condition with some general norms. Towards this, we firstly observe that for the two limiting cases ℓ_1 norm and ℓ_∞ norm, one has $\|p_1 - p_2\|_1 \leq 2$ and $\|p_1 - p_2\|_\infty \leq 1$ for any two probability distribution $p_1, p_2 \in \mathbb{R}^S$. Namely, the accessible ranges of the uncertainty level σ for ℓ_1 norm and ℓ_∞ norm are $(0,2]$ and $(0,1]$, respectively. In addition, we have

$$\forall p_1, p_2 \in \mathbb{R}^S: \quad \|p_1 - p_2\|_\infty \leq \|p_1 - p_2\| \leq \|p_1 - p_2\|_1 \tag{213}$$

for any norm $\|\cdot\|$. It indicates that the accessible range of the uncertainty level $\sigma_{\|\cdot\|}$ for any given norm $\|\cdot\|$ is between $\big(0, \sigma_{\|\cdot\|}^{\max}\big]$, where $1 \leq \sigma_{\|\cdot\|}^{\max} \leq 2$.

To continue, we specify the definition of the uncertainty set with sa-rectangularity condition with some given general norm $\|\cdot\|$ as below: for any nominal transition kernel $P \in \mathbb{R}^{SA\times S}$,

$$\mathcal{U}^{\sigma}_{\|\cdot\|}(P) := \mathcal{U}^{\sigma}_{\|\cdot\|}(P) = \otimes\, \mathcal{U}^{\sigma}_{p}(P_{s,a}), \qquad \mathcal{U}^{\sigma}_{\|\cdot\|}(P_{s,a}) := \Big\{ P'_{s,a} \in \Delta(\mathcal{S}) : \big\|P'_{s,a} - P_{s,a}\big\| \leq \sigma_{\|\cdot\|} \Big\}. \tag{214}$$

Then, we recall the assumption of the uncertainty radius $\sigma_{\|\cdot\|} \in \big(0, \sigma_{\|\cdot\|}^{\max}(1 - c_0)\big]$ with $0 < c_0 < 1$.

Then, resorting to the same class of hard MDPs in [Shi et al., 2023, Section C.1], we can complete the proof by directly following the same proof pipeline of Shi et al. [2023, Section C] by replacing σ with $\sigma_{\|\cdot\|}^{\max}\sigma_{\|\cdot\|}$.

11 Proof of Theorem 4

Developing the lower bound for the cases with s-rectangular uncertainty set involves several new challenges compared to that of (s,a)-rectangular cases. Specifically, the first challenge is that the optimal policy can be stochastic and hard to be characterized with a closed form for the RMDPs with a s-rectangular uncertainty set, rather than deterministic polices in (s,a)-rectangular cases. Such richer and smoother class of optimal policies makes slightly changing the transition kernel generally could only leads to a smoothly changed stochastic optimal policy instead of a completely different one. Such reduced changing of optimal policy further gives smaller performance gap, thus challenges of a tighter lower bound. Second, most of the hard instances in the literature are constructed as SA states with a constant number of action spaces without loss of generality. While when it comes to s-rectangular uncertainty set, the action space size becomes important and can't be assumed as a constant anymore. So a new class of instances are required.

To address these challenges, in this section, we construct a new set of hard RMDP instances for two limiting cases: ℓ_1 norm and ℓ_∞ norm.

11.1 Construction of the hard problem instances

Before proceeding, we introduce two useful sets related to the state space and action space as below:

$$\mathcal{S} = \{0, 1, \ldots, S\}, \qquad \text{and} \qquad \mathcal{A} = \{0, 1, \cdots, A-1\}.$$

In this section, we construct a set of RMDPs termed as $\mathcal{M}_{\ell_\infty}$, which consists of $S(A-1)$ components including $S(A-1)$ components, each associates with some different state-action pair. Specifically, it is defined as

$$\mathcal{M}_{\ell_\infty} := \big\{ \mathcal{M}_\theta = \big(\mathcal{S}, \mathcal{A}, \mathcal{U}^{\sigma}(P^\theta), r, \gamma\big) \mid \theta \in \Theta = \big\{(i,j) : (i,j) \in \mathcal{S} \times \mathcal{A} \setminus \{0\}\big\} \big\}. \tag{215}$$

We introduce the detailed definition of $\mathcal{M}_{\ell_\infty}$ by introducing several key components of it sequentially. In particular, for any RMDP $\mathcal{M}_\theta \in \mathcal{M}_{\ell_\infty}$, the state space is of size $2S$, which includes two classes of states $\mathcal{X} = \{x_0, x_1, \cdots, x_{S-1}\}$ and $\mathcal{Y} = \{y_0, y_1, \cdots, y_{S-1}\}$. The action space for each state is $\mathcal{A}$ of A possible actions. So we have totally $2S$ states and there is in total $2SA$ state-action pairs.

Armed with the above definitions, we can first introduce the following nominal transition kernel: for all $(s,a) \in \mathcal{X} \cup \mathcal{Y} \times \mathcal{A}$

$$P^{(0,0)}(s' \mid s,a) = \begin{cases} p\mathbb{1}(s'=y_i) + (1-p)\mathbb{1}(s'=x_i) & \text{if} \quad s = x_i, a = 0, \quad \forall i \in \mathcal{S} \\ q\mathbb{1}(s'=y_i) + (1-q)\mathbb{1}(s'=x_i) & \text{if} \quad s = x_i, a \neq 0, \quad \forall i \in \mathcal{S} \\ \mathbb{1}(s'=s) & \text{if} \quad s \in \mathcal{Y} \end{cases} \tag{216}$$

Here, p and q are set according to

$$0 \leq p \leq 1 \quad \text{and} \quad 0 \leq q = p - \Delta \tag{217}$$

for some p and $\Delta > 0$ that will be introduced momentarily.

Then we introduce the $S(A-1)$ components inside $\mathcal{M}_\infty$. Namely, for any $(i,j) \in \mathcal{S} \times \mathcal{A} \setminus \{0\}$, the nominal transition kernel of $\mathcal{M}_{(i,j)}$ is specified as

$$P^{(i,j)}(s' \mid s,a) = \begin{cases} p\mathbb{1}(s'=y_i) + (1-p)\mathbb{1}(s'=x_i) & \text{if } s = x_i, a = j \\ q\mathbb{1}(s'=y_i) + (1-q)\mathbb{1}(s'=x_i) & \text{if } s = x_i \in \mathcal{X}, a = 0 \\ P^{(0,0)}(s' \mid s,a) & \text{otherwise} \end{cases} \tag{218}$$

In words, the nominal transition kernel of each variant $\mathcal{M}_{(i,j)}$ only differs slightly from that of the basic nominal transition kernel $P^{(0,0)}$ when $s = x_i$ and $a = \{0, j\}$, which makes all the components inside $\mathcal{M}_{\ell_\infty}$ closed to each other.

In addition, the reward function is defined as

$$\forall a \in \mathcal{A}: \quad r(s,a) = \begin{cases} 1 & \text{if } s \in \mathcal{Y} \\ 0 & \text{otherwise.} \end{cases} \tag{219}$$

Uncertainty set of the transition kernels. Recall the following useful notation for any transition probability P, i.e., the transition vector associated with some state s is denoted as:

$$P_s := P(\cdot,\cdot \mid s) \in \mathbb{R}^{1 \times SA}, \quad P_s^0 := P^0(\cdot,\cdot \mid s) \in \mathbb{R}^{1 \times SA}. \tag{220}$$

With this in hand, the uncertainty set (definition in (5)) with ℓ_∞ norm for any P^θ with $\theta \in \Theta$ can be represented as:

$$\mathcal{U}_\infty^{\mathsf{s},\widetilde{\sigma}}(P_s^\theta) := \mathcal{U}_{\|\cdot\|}^{\mathsf{s},\widetilde{\sigma}}(P_s^\theta) = \Big\{ P_s' \in \Delta(\mathcal{S})^{\mathcal{A}} : \big\| P_s' - P_s^\theta \big\| \leq \tilde{\sigma} = \sigma \left\| 1 \right\|_\infty = \sigma \Big\}. \tag{221}$$

So without loss of generality, we set the radius $\sigma \in (0, (1-c_0)]$ with $0 < c_0 < 1$. Before proceeding, we observe that as the uncertainty set above is defined with respect to ℓ_∞, it directly implies that for each $(s,a) \in \mathcal{S} \times \mathcal{A}$, the uncertainty set is independent and can be decomposed as

$$\mathcal{U}_\infty^{\mathsf{s},\widetilde{\sigma}}(P_s^\theta) = \otimes \mathcal{U}_{\|\cdot\|}^{\mathsf{s},\widetilde{\sigma}}(P_{s,a}^\theta) = \Big\{ P_{s,a}' \in \Delta(\mathcal{S}) : \big\| P_{s,a}' - P_{s,a}^\theta \big\| \leq \sigma \Big\}. \tag{222}$$

Notably, this indicates that using s-rectangular uncertainty set with ℓ_∞ norm as the divergence function is analogous to the case of using (s,a)-rectangular uncertainty set with ℓ_∞ norm. As a result, we follow the pipeline of the prior art Shi et al. [2023, Section C] which established the minimax-optimal lower bound for (s,a)-rectangular RMDPs with TV distance, which is analogous to the ℓ_∞ case. Towards this, we set p, q, Δ as the same as the ones in Shi et al. [2023, Section C.1], where we recall the expressions of p, q, Δ for self-contained as below: taking $c_1 := \frac{c_0}{2}$,

$$p = (1 + c_1) \max\{1-\gamma, \sigma\} \qquad \text{and} \qquad \Delta \leq c_1 \max\{1-\gamma, \sigma\}, \tag{223}$$

which ensure several facts:

$$0 \leq p \leq 1 \quad \text{and} \quad p \geq q \geq \max\{1-\gamma, \sigma\}. \tag{224}$$

Value functions and optimal policies. For each RMDP instance $\mathcal{M}_\theta \in \mathcal{M}_{\ell_\infty}$, with some abuse of notation, we denote $\pi_\theta^\star$ as the optimal policy. In addition, let $V_\theta^{\pi,\sigma}$ (resp. $V_\theta^{\star,\sigma}$) represent the corresponding robust value function of any policy π (resp. $\pi_\theta^\star$) with uncertainty level σ. Armed with these notations, the following lemma shows some essential properties concerning the value functions and optimal policies; the proof is postponed to Appendix 11.3.

Lemma 12. *Consider any $\mathcal{M}_\theta \in \mathcal{M}_{\ell_\infty}$ and any policy π, one has*

$$\forall (i,j) \in \Theta : \quad V_{(i,j)}^{\pi,\sigma}(x_i) \leq \frac{\gamma\big(z_{(i,j)}^\pi - \sigma\big)}{(1-\gamma)\left(1 + \frac{\gamma\left(z_{(i,j)}^\pi - \sigma\right)}{1-\gamma(1-\sigma)}\right)(1-\gamma(1-\sigma))}, \tag{225}$$

where $z_{(i,j)}^\pi$ is defined as

$$\forall (i,j) \in \Theta : \quad z_{(i,j)}^\pi \coloneqq p\pi(j \mid x_i) + q\left[1 - \pi(j \mid x_i)\right]. \tag{226}$$

In addition, the robust optimal value functions and the robust optimal policies satisfy

$$\forall (i,j) \in \Theta, s \in \mathcal{X} : \quad V_{(i,j)}^{\star,\sigma}(s) = \frac{\gamma(p-\sigma)}{(1-\gamma)\left(1 + \frac{\gamma(p-\sigma)}{1-\gamma(1-\sigma)}\right)(1-\gamma(1-\sigma))} \tag{227}$$

and

$$\pi_{(i,j)}^\star(j \mid x_i) = 1 \qquad \textit{and} \qquad \pi_{(i,j)}^\star(0 \mid s) = 1 \quad \forall s \in \mathcal{X} \setminus \{x_i\}. \tag{228}$$

In words, this lemma shows that for any RMDP $\mathcal{M}_{(i,j)}$, the optimal policy on state x_i satisfies $\pi_{(i,j)}^\star(j \mid x_i) = 1$ and will focus on $a = 0$ for all other states $s \in \mathcal{X} \setminus \{x_i\}$.

11.2 Establishing the minimax lower bound

Step 1: converting the goal to estimate (i,j). Now we are in position to derive the lower bound. Recall the goal is to control the following quantity associated with any policy estimator $\widehat{\pi}$ based on the dataset with in total N_{all} samples:

$$\max_{(i,j)\in\Theta} \mathbb{P}_{(i,j)}\left\{\max_{s\in\mathcal{X}\cup\mathcal{Y}}\left(V_{(i,j)}^{\star,\sigma}(s) - V_{(i,j)}^{\widehat{\pi},\sigma}(s)\right)\right\} \geq \max_{(i,j)\in\Theta} \mathbb{P}_{(i,j)}\left\{\max_{s\in\mathcal{X}}\left(V_{(i,j)}^{\star,\sigma}(s) - V_{(i,j)}^{\widehat{\pi},\sigma}(s)\right)\right\}. \tag{229}$$

To do so, we can invoke a key claim in Shi et al. [2023] here since our problem setting can be reduced to the same one in Shi et al. [2023]: With $\varepsilon \leq \frac{c_1}{32(1-\gamma)}$, letting

$$\Delta = 32(1-\gamma)\max\{1-\gamma, \sigma\}\varepsilon \leq c_1 \max\{1-\gamma, \sigma\} \tag{230}$$

which satisfies (223), it leads to that for any policy $\widehat{\pi}$ and all $(i,j) \in \Theta$,

$$\begin{aligned} V_{(i,j)}^{\star,\sigma}(x_i) - V_{(i,j)}^{\widehat{\pi},\sigma}(x_i) &\geq 2\varepsilon\big(1 - \widehat{\pi}(j \mid x_i)\big), \\ \forall s \in \mathcal{X} \setminus \{x_i\} : \quad V_{(i,j)}^{\star,\sigma}(s) - V_{(i,j)}^{\widehat{\pi},\sigma}(s) &\geq 2\varepsilon\big(1 - \widehat{\pi}(0 \mid s)\big). \end{aligned} \tag{231}$$

Before continuing, we introduce a useful notation for the subset of Θ excluding the cases with state i is selected:

$$\forall i \in \mathcal{S} : \quad \Theta_{-i} = \Theta \setminus \{(i',j) : i' = i, j \in \mathcal{A} \setminus \{0\}\}. \tag{232}$$

Armed with the above facts and notations, we first suppose there exists a policy $\widehat{\pi}$ such that for some $(i,j) \in \Theta$,

$$\mathbb{P}_{(i,j)}\left\{V_{(i,j)}^{\star,\sigma}(x_i) - V_{(i,j)}^{\widehat{\pi},\sigma}(x_i) \leq \varepsilon\right\} \geq \frac{3}{4}. \tag{233}$$

which in view of (231) indicates that we necessarily have $\widehat{\pi}(j \mid x_i) \geq \frac{1}{A}$ with probability at least $\frac{3}{4}$.

As a result, taking

$$j' = \arg\max_{a\in\mathcal{A}} \widehat{\pi}(a \,|\, x_i), \tag{234}$$

we are motivated to construct the following estimate of θ:

$$\widehat{\theta} \begin{cases} = (i, j') & \text{if} \quad j' > 0 \\ \in \mathcal{G}_{-w} & \text{if} \quad j' = 0, \end{cases} \tag{235}$$

which satisfies

$$\mathbb{P}_{(i,j)}\big\{\widehat{\theta} = (i,j)\big\} \geq \mathbb{P}_{(i,j)}\big\{j' = j\big\} \geq \mathbb{P}_{(i,j)}\big\{\widehat{\pi}(j \,|\, x_i) > \frac{1}{A}\big\} \geq \frac{3}{4}. \tag{236}$$

Step 2: developing the probability of error in testing multiple hypotheses. Before proceeding, we discuss the dataset consisting of in total N_{all} independent samples. Observing that each RMDP inside the set $\mathcal{M}_{\ell_\infty}$ are constructed symmetrically associated with one pair of states (x_i, y_i) for all $i \in \mathcal{S}$ and another action $j \in \mathcal{A} \times \{0\}$, respectively. Therefore, it is obvious that the dataset is supposed to be generated uniformly on each (x_i, y_i, j) to maximize the information gain, leading to $\frac{N_{\mathsf{all}}}{S(A-1)}$ samples for any states-action (x_i, y_i, j) with $i \in \mathcal{S}, j \in \mathcal{A} \setminus \{0\}$.

Then we are ready to turn to the hypothesis testing problem over $(i,j) \in \Theta$. Towards this, we consider the minimax probability of error defined as follows:

$$p_{\mathrm{e}} := \inf_{\phi} \max_{(i,j)\in\Theta} \big\{\mathbb{P}_{(i,j)}\big(\phi \neq (i,j)\big)\big\}, \tag{237}$$

where the infimum is taken over all possible tests ϕ constructed from the dataset introduced above.

To continue, armed with the above dataset with N_{all} independent samples, we denote $\mu^{i,j}$ (resp. $\mu^{i,j}(s,a)$) as the distribution vector (resp. distribution) of each sample tuple (s,a,s') under the nominal transition kernel $P^{(i,j)}$ associated with $\mathcal{M}_{(i,j)}$. With this in mind, combined with Fano's inequality from Tsybakov [2009, Theorem 2.2] and the additivity of the KL divergence (cf. Tsybakov [2009, Page 85]), we obtain

$$\begin{aligned}
p_{\mathrm{e}} &\geq 1 - N_{\mathsf{all}} \frac{\max_{(i,j),(i',j')\in\Theta,(i,j)\neq(i',j')} \mathsf{KL}\big(\mu^{i,j} \,|\, \mu^{i',j'}\big) + \log 2}{\log|\Theta|} \\
&\overset{\text{(i)}}{\geq} 1 - N_{\mathsf{all}} \max_{(i,j),(i',j')\in\Theta,(i,j)\neq(i',j')} \mathsf{KL}\big(\mu^{i,j} \,|\, \mu^{i',j'}\big) - \frac{1}{2} \\
&= \frac{1}{2} - N_{\mathsf{all}} \max_{(i,j),(i',j')\in\Theta,(i,j)\neq(i',j')} \mathsf{KL}\big(\mu^{i,j} \,|\, \mu^{i',j'}\big)
\end{aligned} \tag{238}$$

where (i) holds by $\log|\Theta| \geq 2\log 2$ as long as $S(A-1)$ are large enough.

Then following the same proof pipeline of Shi et al. [2023, Section C.2], we can arrive at

$$p_{\mathrm{e}} \geq \frac{1}{2} - \frac{N_{\mathsf{all}}}{S(A-1)} \frac{4096}{c_1} (1-\gamma)^2 \max\{1-\gamma, \sigma\}\varepsilon^2 \geq \frac{1}{4}, \tag{239}$$

if the sample size is selected as

$$N_{\mathsf{all}} \leq \frac{c_1 S(A-1)}{16396(1-\gamma)^2 \max\{1-\gamma, \sigma\}\varepsilon^2}. \tag{240}$$

Step 3: summing up the results together. Finally, we suppose that there exists an estimator $\widehat{\pi}$ such that

$$\max_{(i,j)\in\Theta} \mathbb{P}_{(i,j)}\left[\max_{s\in\mathcal{X}\cup\mathcal{Y}} \left(V^{\star,\sigma}_{(i,j)}(s) - V^{\widehat{\pi},\sigma}_{(i,j)}(s)\right) \geq \varepsilon\right] < \frac{1}{4}, \tag{241}$$

then according to (229), we necessarily have

$$\forall s \in \mathcal{X}: \quad \max_{(i,j)\in\Theta} \mathbb{P}_{(i,j)}\left[V^{\star,\sigma}_{(i,j)}(s) - V^{\widehat{\pi},\sigma}_{(i,j)}(s) \geq \varepsilon\right] < \frac{1}{4}, \tag{242}$$

which indicates

$$\forall s \in \mathcal{X}: \quad \max_{(i,j)\in\Theta} \mathbb{P}_{(i,j)} \left[V^{\star,\sigma}_{(i,j)}(s) - V^{\widehat{\pi},\sigma}_{(i,j)}(s) < \varepsilon \right] \geq \frac{3}{4}. \tag{243}$$

As a consequence, (236) shows we must have

$$\forall (i,j) \in \Theta: \quad \mathbb{P}_{(i,j)} \left[\widehat{\theta} = (i,j) \right] \geq \frac{3}{4} \tag{244}$$

to achieve (241). However, this would contract with (239) if the sample size condition in (240) is satisfied. Thus, we complete the proof.

11.3 Proof of Lemma 12

Without loss of generality, we first consider any $\mathcal{M}_{(i,j)}$ with $(i,j) \in \mathcal{S} \times \mathcal{A} \setminus \{0\}$. Following the same routine of Shi et al. [2023, Section C.3.1], we can verify that the order of the robust value function $V^{\pi,\sigma}_{(i,j)}$ over different states satisfies

$$\forall k \in \mathcal{S}: \quad V^{\pi,\sigma}_{(i,j)}(x_k) \leq V^{\pi,\sigma}_{(i,j)}(y_k), \tag{245}$$

which means the robust value function of the states inside $\mathcal{X}$ are always not larger than the corresponding states inside $\mathcal{Y}$.

Then we denote the minimum of the robust value function over states as below:

$$V^{\pi,\sigma}_{(i,j),\min} \coloneqq \min_{s\in\mathcal{S}} V^{\pi,\sigma}_{(i,j)}(s). \tag{246}$$

In the following arguments, we first take a moment to assume $V^{\pi,\sigma}_{(i,j),\min} = V^{\pi,\sigma}_{(i,j)}(x_i)$. With this in mind, we arrive at

$$V^{\pi,\sigma}_{(i,j)}(y_i) = 1 + \gamma(1-\sigma) V^{\pi,\sigma}_{(i,j)}(y_i) + \gamma\sigma V^{\pi,\sigma}_{(i,j),\min} = \frac{1 + \gamma\sigma V^{\pi,\sigma}_{(i,j)}(x_i)}{1 - \gamma(1-\sigma)}. \tag{247}$$

Then, when we move on to the characterization of the robust value function at state x_i. To do so, we notice two important facts:

1) The nominal transition probability $P^{(i,j)}_{x_i,a}$ at state-action pair (x_i, a) for any $a \in \mathcal{A}$ is a Bernoulli distribution (see (218) and (216)). The TV distance and the ℓ_∞ norm between two Bernoulli distribution are the same.
2) Invoking the definitions of the nominal transition probability in (218) and (216), we have

$$\begin{aligned} P^{(i,j)}_{x_i,j} &= p\mathbb{1}(s' = y_i) + (1-p)\mathbb{1}(s' = x_i) \\ P^{(i,j)}_{x_i,a} &= q\mathbb{1}(s' = y_i) + (1-q)\mathbb{1}(s' = x_i) \quad \forall a \in \mathcal{A} \setminus \{j\}. \end{aligned} \tag{248}$$

With the above two facts in hand, our problem setting is reduced to the same one in Shi et al. [2023] and can reuse the results in Shi et al. [2023, Section C.3.1] to achieve

$$V^{\pi,\sigma}_{(i,j)}(x_i) \leq \frac{\frac{\gamma\left(z^{\pi}_{(i,j)} - \sigma\right)}{1-\gamma(1-\sigma)}}{(1-\gamma)\left(1 + \frac{\gamma\left(z^{\pi}_{(i,j)} - \sigma\right)}{1-\gamma(1-\sigma)}\right)}. \tag{249}$$

and

$$\begin{aligned} &\pi^\star_{(i,j)}(j \,|\, x_i) = 1 \\ &V^{\star,\sigma}_{(i,j)}(x_i) = \frac{\frac{\gamma\left(z^{\pi^\star}_{(i,j)} - \sigma\right)}{1-\gamma(1-\sigma)}}{(1-\gamma)\left(1 + \frac{\gamma\left(z^{\pi^\star}_{(i,j)} - \sigma\right)}{1-\gamma(1-\sigma)}\right)} = \frac{\frac{\gamma(p-\sigma)}{1-\gamma(1-\sigma)}}{(1-\gamma)\left(1 + \frac{\gamma(p-\sigma)}{1-\gamma(1-\sigma)}\right)}. \end{aligned} \tag{250}$$

Analogously, we can verify that for other $x_k \in \mathcal{X} \setminus \{x_i\}$,

$$\begin{aligned} &\pi^\star_{(i,j)}(0 \,|\, x_k) = 1 \\ &V^{\star,\sigma}_{(i,j)}(x_k) = \frac{\frac{\gamma(p-\sigma)}{1-\gamma(1-\sigma)}}{(1-\gamma)\left(1 + \frac{\gamma(p-\sigma)}{1-\gamma(1-\sigma)}\right)}. \end{aligned} \tag{251}$$

12 DRVI for $sa-$ rectangular algorithm for arbitrary norm

In order to compute the fixed point of $\widehat{\mathcal{T}}^{\sigma}$, distributionally robust value iteration (DRVI), is defined in Algorithm 1. For sa-rectangularity, starting from an initialization $\widehat{Q}_0 = 0$, the update rule at the t-th ($t \geq 1$) iteration is the following $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$:

$$\widehat{Q}_t^{\pi}(s,a) = \widehat{\mathcal{T}}^{\sigma} \widehat{Q}_{t-1}^{\pi}(s,a) = r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\mathsf{sa},\sigma}_{\|\cdot\|}(\widehat{P}^0_{s,a})} \mathcal{P} \widehat{V}_{t-1}, \tag{252}$$

where $\widehat{V}_{t-1}(s) = \max_{\pi} \widehat{Q}_{t-1}^{\pi}(s,a)$ for all $s \in \mathcal{S}$.

Directly solving (252) is computationally expensive since it involves optimization over a S-dimensional probability simplex at each iteration, especially when the dimension of the state space $\mathcal{S}$ is large. Fortunately, given strong duality (252) can be equivalently solved using its dual problem, which concerns optimizing a two variable (λ and ω) and thus can be solved efficiently. The specific form of the dual problem depends on the choice of the norm $\|.\|$, which we shall discuss separately in Appendix 8.3. To complete the description, we output the greedy policy of the final Q-estimate $\widehat{Q}_T$ as the final policy $\widehat{\pi}$, namely,

$$\forall s \in \mathcal{S}: \quad \widehat{\pi}(s) = \arg\max_{a} \widehat{Q}_T(s,a). \tag{253}$$

Encouragingly, the iterates $\left\{\widehat{Q}_t\right\}_{t \geq 0}$ of $DRVI$ converge linearly to the fixed point $\widehat{Q}^{\star,\sigma}$, owing to the appealing γ-contraction property of $\widehat{\mathcal{T}}^{\sigma}$.

input: empirical nominal transition kernel $\widehat{P}^0$; reward function r; uncertainty level σ; number of iterations T.
initialization: $\widehat{Q}_0(s,a) = 0$, $\widehat{V}_0(s) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.
for $t = 1, 2, ..., T$ **do**
  for $s \in \mathcal{S}, a \in \mathcal{A}$ **do**
    Set $\widehat{Q}_t(s,a)$ according to (252);
  end
  for $s \in \mathcal{S}$ **do**
    Set $\widehat{V}_t(s) = \max_a \widehat{Q}_t(s,a)$;
  end
end
output: $\widehat{Q}_T$, $\widehat{V}_T$ and $\widehat{\pi}$ obeying $\widehat{\pi}(s) := \arg\max_a \widehat{Q}_T(s,a)$.

Algorithm 1: Distributionally robust value iteration ($DRVI$) for infinite-horizon RMDPs for sa-rectangular for arbitrary norm

Using Algorithm 1, it allows getting an ϵ_{opt} error in the empirical MDP in the sa-rectangular case. In the s-rectangular case, finding an algorithm to get ϵ_{opt} is more difficult to use, as the policy is not deterministic anymore and 1 cannot anymore be applied. For L_p norms, Clavier et al. [2023] derived an algorithm but for arbitrary norm we need to consider a more general problem for arbitrary norm in Appendix 12

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See conclusion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Assumtions are stated in lemmas and Theorems.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: Theoretical paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: Theoretical paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: Theoretical paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Theoretical paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: Theoretical paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:Done

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification:Theoretical paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Theoretical paper.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: Theoretical paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification:Theoretical paper.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Theoretical paper.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Theoretical paper.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.