Online Policy Optimization for Robust MDP

Jing Dong *
The Chinese University of Hong Kong, Shenzhen
jingdong@link.cuhk.edu.cn

Jingwei Li *
Tsinghua University
ljw22@mails.tsinghua.edu.cn

Baoxiang Wang *
The Chinese University of Hong Kong, Shenzhen
bxiangwang@cuhk.edu.cn

Jingzhao Zhang *†
Tsinghua University
jingzhaoz@mail.tsinghua.edu.cn

Abstract

Reinforcement learning (RL) has exceeded human performance in many synthetic settings such as video games and Go. However, real-world deployment of end-to-end RL models is less common, as RL models can be very sensitive to slight perturbation of the environment. The robust Markov decision process (MDP) framework—in which the transition probabilities belong to an uncertainty set around a nominal model—provides one way to develop robust models. While previous analysis shows RL algorithms are effective assuming access to a generative model, it remains unclear whether RL can be efficient under a more realistic online setting, which requires a careful balance between exploration and exploitation. In this work, we consider online robust MDP by interacting with an unknown nominal system. We propose a robust optimistic policy optimization algorithm that is provably efficient. To address the additional uncertainty caused by an adversarial environment, our model features a new optimistic update rule derived via Fenchel conjugates. Our analysis establishes the first regret bound for online robust MDPs.

1 Introduction

The rapid progress of reinforcement learning (RL) algorithms enables trained agents to navigate around complicated environments and solve complex tasks. The standard reinforcement learning methods, however, may fail catastrophically in another environment, even if the two environments only differ slightly in dynamics [11, 22, 7, 31, 25]. In practical applications, such mismatch of environment dynamics are common and can be caused by a number of reasons, e.g., model deviation due to incomplete data, unexpected perturbation and possible adversarial attacks. To model the potential mismatch between system dynamics, the framework of robust MDP is introduced to account for the uncertainty of the parameters of the MDP [27, 35, 21, 12]. Under this framework, the dynamic of an MDP is no longer fixed but can come from some uncertainty set, such as the rectangular uncertainty set, centered around a nominal transition kernel. The agent sequentially interacts with the nominal transition kernel to learn a policy, which is then evaluated on the worst possible transition from the uncertainty set. Therefore, the objective is to find the worst-case best-performing policy.

If a generative model (also known as a simulator) of the environment or a suitable offline dataset is available, one could obtain a ϵ-optimal robust policy with $\tilde{O}(\epsilon^{-2})$ samples under a rectangular uncertainty set [24, 23, 34, 18]. Yet the presence of a generative model is stringent to fulfill for real

*Author names are listed in alphabetical order.

†Jingzhao Zhang is also affiliated with Shanghai Qi Zhi Institute and Shanghai Artificial Intelligence Laboratory.

Workshop on Robustness in Sequence Modeling, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

applications. In a more practical online setting, the agent sequentially interacts with the environment and tackles the exploration-exploitation challenge as it balances between exploring the state space and exploiting the high-reward actions. in the online setting, which is captured by the regret, is more challenging to achieve than algorithm convergence. In the robust MDP setting, previous sample complexity results cannot directly imply a sublinear regret in general Dann et al. [8] and so far no asymptotic result is available. A more detailed review of the related works are deferred to the Appendix.

In this paper, we propose the first policy optimization algorithm for robust MDP under a rectangular uncertainty set. One of the challenges for deriving a regret guarantee for robust MDP stems from its adversarial nature. As the transition dynamic can be picked adversarially from a predefined set, the optimal policy is in general randomized [36]. This is in contrast with conventional MDPs, where there always exists a deterministic optimal policy, which can be found with value-based methods and a greedy policy (e.g. UCB-VI algorithms). Bearing this observation, we resort to policy optimization (PO)-based methods, which directly optimize a stochastic policy in an incremental way.

With a stochastic policy, our algorithm explores robust MDPs in an optimistic manner. To achieve this robustly, we propose a carefully designed bonus function via the dual conjugate of the robust bellman equation. This quantifies both the uncertainty stemming from the limited historical data and the uncertainty of the MDP dynamic. In the episodic setting of robust MDPs, we show that our algorithm attains sublinear regret $O(\sqrt{K})$ for both (s, a) and s-rectangular uncertainty set, where K is the number of episodes. In the case where the uncertainty set contains only the nominal transition model, our results recover the previous regret upper bound of non-robust policy optimization [30]. Our result achieves the first provably efficient regret bound in the online robust MDP problem. We further validated our algorithm with experiments.

2 Problem formulation

In this section, we describe the formal setup of robust MDP. We start with defining some notations.

Robust Markov decision process We consider an episodic finite horizon robust MDP, which can denoted by a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, H, \{\mathcal{P}\}_{h=1}^{H}, \{r\}_{h=1}^{H} \rangle$. Here $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\{r\}_{h=1}^{H}$ is the time-dependent reward function, and H is the length of each episode. Instead of a fixed step of time-dependent uncertainty kernels, the transitions of the robust MDP is governed by kernels that are within a time-dependent uncertainty set $\{\mathcal{P}\}_{h=1}^{H}$, $i.e.$, time-dependent transition $P_h \in \mathcal{P}_h \subseteq \Delta_{\mathcal{S}}$ at time h. We consider the case where the rewards are stochastic. This is, on state-action (s, a) at time h, the immediate reward is $R_h(s, a) \in [0, 1]$, which is drawn i.i.d from a distribution with expectation $r_h(s, a)$. With the described setup of robust MDPs, we now define the policy and its associated value.

Policy and robust value function A time-dependent policy π is defined as $\pi = \{\pi_h\}_{h=1}^{H}$, where each π_h is a function from $\mathcal{S}$ to the probability simplex over actions, $\Delta(\mathcal{A})$. If the transition kernel is fixed to be P, the performance of a policy π starting from state s at time h can be measured by its value function, which is defined as $V_h^{\pi,P}(s) = \mathbb{E}_{\pi,P}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \mid s_h = s\right]$. In robust MDP, the robust value function instead measures the performance of π under the worst possible choice of transition P within the uncertainty set. Specifically, the value and the Q-value function of a policy given the state action pair (s, a) at step h are defined as

$$V_h^{\pi}(s) = \min_{\{P_h\} \in \{\mathcal{P}_h\}} V_h^{\pi,\{P\}}(s)\,,$$

$$Q_h^{\pi}(s, a) = \min_{\{P_h\} \in \{\mathcal{P}_h\}} \mathbb{E}_{\pi,\{P\}}\left[\sum_{h'=h}^{H} r_h(s_{h'}, a_{h'}) \mid (s_h, a_h) = (s, a)\right].$$

The optimal value function is defined to be the best possible value attained by a policy $V_h^*(s) = \max_\pi V_h^\pi(s) = \max_\pi \min_{\{P_h\} \in \{\mathcal{P}_h\}} V_h^{\pi,\{P\}}(s)$. The optimal policy is then defined to be the policy that attains the optimal value.

Robust Bellman equation Similar to non-robust MDP, robust MDP has the following robust bellman equation, which characterizes a relation to the robust value function. $Q_h^{\pi}(s, a) =$

$r(s,a)+\sigma_{\mathcal{P}_h}(V^\pi_{h+1})(s,a)\,, \quad V^\pi_h(s) = \langle Q^\pi_h(s,\cdot), \pi_h(\cdot,s)\rangle$, where $\sigma_{\mathcal{P}_h}(V^\pi_{h+1})(s,a) = \min_{P_h\in\mathcal{P}_h} P_h(\cdot \mid s,a)V^\pi_{h+1}\,, P_h(\cdot \mid s,a)V = \sum_{s'\in\mathcal{S}} P_h(s' \mid s,a)V(s')\,.$

Without additional assumptions on the uncertainty set, the optimal policy and value of the robust MDP are in general NP-hard to solve [36]. Thus, to limit the level of perturbations, we assume that the transition kernels is close to the nominal transition measured via ℓ_1 distance. We consider two cases.

Definition 2.1 ((s,a)-rectangular uncertainty set Iyengar [12], Wiesemann et al. [36])**.** *For all time step h and with a given state-action pair (s,a), the (s,a)-rectangular uncertainty set $\mathcal{P}_h(s,a)$ is defined as $\mathcal{P}_h(s,a) = \left\{\|P_h(\cdot \mid s,a) - P^o_h(\cdot \mid s,a)\|_1 \le \rho, P_h(\cdot \mid s,a) \in \Delta(\mathcal{S})\right\}$, where P^o_h is the nominal transition kernel at h, $P^o_h(\cdot \mid s,a) > 0, \forall (s,a) \in \mathcal{S}\times\mathcal{A}$, ρ is the level of uncertainty.*

One way to relax the (s,a)-rectangular assumption is to instead let the uncertain transition kernels within the set take value independent for each s only. This characterization is then more general and its solution gives a stronger robustness guarantee.

Definition 2.2 (s-rectangular uncertainty set Wiesemann et al. [36])**.** *For all time step h and with a given state s, the s-rectangular uncertainty set $\mathcal{P}_h(s)$ is defined as $\mathcal{P}_h(s) = \left\{\sum_{a\in\mathcal{A}} \|P_h(\cdot \mid s,a) - P^o_h(\cdot \mid s,a)\|_1 \le A\rho, P_h(\cdot \mid s,\cdot) \in \Delta(\mathcal{S})^{\mathcal{A}}\right\}$, where P^o_h is the nominal transition kernel at h, $P^o_h(\cdot \mid s,a) > 0, \forall (s,a) \in \mathcal{S}\times\mathcal{A}$, ρ is the level of uncertainty.*

Different from the (s,a)-rectangular assumption, which guarantees the existence of a deterministic optimal policy, the optimal policy under s-rectangular set may need to be randomized [36]. We also remark that the requirement of $P^o_h(\cdot \mid s,a) > 0$ is mostly for technical convenience.

Equipped with the characterization of the uncertainty set, we now describe the definition of regret under the robust MDP.

Learning protocols and regret We consider a learning agent repeatedly interacts with the environment defined by the nominal transition model in an episodic manner, over K episodes. We remark that if the agent is asked to interact with a potentially adversarially chosen transition, the learning problem is NP-hard [10]. We assume the agents always start from a fixed initial state s. The performance of the learning agent is measured by the cumulative regret incurred over the K episodes, which is defined to be the cumulative difference between the robust value of π_k and the robust value of the optimal policy. That is, $\sum_{k=1}^K V^*_1(s_0) - V^{\pi_k}_1(s_0)$, where s^k_0 is the initial state.

3 Algorithm

Our algorithm performs policy optimization with empirical estimates and encourages exploration by adding a bonus to less explored states. However, we need to propose a new efficiently computable bonus that is robust to adversarial transitions. We achieve this via solving a sub-optimization problem derived from Fenchel conjugate. We present Robust Optimistic Policy Optimization (ROPO) and elaborate on its design components.

To start, as our algorithm has no access to the actual reward and transition function, we use the following empirical estimator of the transition and reward:

$$\hat{r}^k_h(s,a) = \frac{\sum_{k'=1}^{k-1} R^{k'}_h(s,a)\mathbb{I}\left\{s^{k'}_h = s, a^{k'}_h = a\right\}}{N^k_h(s,a)}\,,$$
$$\hat{P}^{o,k}_h(s,a) = \frac{\sum_{k'=1}^{k-1} \mathbb{I}\left\{s^{k'}_h = s, a^{k'}_h = a, s^{k'}_{h+1} = s'\right\}}{N^k_h(s,a)}\,, \tag{1}$$

where $N^k_h(s,a) = \max\left\{\sum_{k'=1}^{k-1} \mathbb{I}\left\{s^{k'}_h = s, a^{k'}_h = a\right\}, 1\right\}$.

Robust Policy Evaluation step In each episode, the algorithm estimates Q-values with an optimistic variant of the bellman equation. Specifically, to encourage exploration in the robust MDP, we

add a bonus term $b_h^k(s,a)$, which compensates for the lack of knowledge of the actual reward and transition model as well as the uncertainly set, with order $b_h^k(s,a) = O\left(1/\sqrt{N_h^k(s,a)}\right)$.

$$\hat{Q}_h^k(s,a) = \min\left\{\hat{r}(s,a) + \sigma_{\hat{\mathcal{P}}_h}(\hat{V}_{h+1}^{\pi})(s) + b_h^k(s,a), H\right\}.$$

Intuitively, the bonus term b_h^k desires to characterize the optimism required for efficient exploration for both the estimation errors of P and the robustness of P. It is hard to control the two quantities in their primal form because of the coupling between them. We propose the following procedure to address the problem.

Note that the key difference between our algorithm and standard policy optimization is that $\sigma_{\hat{\mathcal{P}}_h}(\hat{V}_{h+1}^{\pi})(s)$ requires solving an inner minimization. Through relaxing the constraints with Lagrangian multiplier and Fenchel conjugates, under (s,a)-rectangular set, the inner minimization problem can be reduced to a one-dimensional unconstrained convex optimization problem on $\mathbb{R}$ (Lemma 4).

$$\sup_{\eta} \eta - \frac{(\eta - \min_s \hat{V}_{h+1}^{\pi_k}(s))_+}{2}\rho - \sum_{s'} \hat{P}_h^o(s' \mid s,a)\left(\eta - \hat{V}_{h+1}^{\pi_k}(s')\right)_+ . \tag{2}$$

The optimum of Equation (2) is then computed efficiently with bisection or sub-gradient methods. Similarly, in the case of s-rectangular set, the inner minimization problem is equivalent to a A-dimensional convex optimization problem, which can be computed efficiently in $\tilde{O}(A)$ iterations by methods like gradient descent. In addition to reducing computational complexity, the dual form decouples the uncertainty in estimation error and in robustness, as ρ and $\hat{P}_h^o$ are not in different terms. The exact form of b_h^k is presented in the Equation (4) and (5). In the case of s-rectangular set, the inner minimization problem is similarly equivalent to the following A-dimensional convex optimization problem.

$$\sup_{\eta} \sum_{a'} \eta_{a'} - \sum_{s',a'} \hat{P}_h^o(s' \mid s,a')\left(\eta_{a'} - \mathbb{I}\{a'=a\}V_{h+1}^{\pi_k}(s')\right)_+ - \min_{s',a'} \frac{A\rho(\eta_{a'} - \mathbb{I}\{a'=a\}V_{h+1}^{\pi_k}(s'))_+}{2}. \tag{3}$$

Policy Improvement Step Using the optimistic Q-value obtained from policy evaluation, the algorithm improves the policy with a KL regularized online mirror descent step,

$$\pi_h^{k+1} \in \arg\max_{\pi} \beta\langle \nabla \hat{V}_h^{\pi_k}, \pi\rangle - \pi_h^k + D_{KL}(\pi||\pi_h^k),$$

where β is the learning rate. In the non-robust case, this improvement step is also shown to be theoretically efficient [30, 37]. Many empirically successful policy optimization algorithms, such as PPO [29] and TRPO [28], also take a similar approach to KL regularization for non-robust policy improvement.

4 Main results

We are now ready to analyze the theoretical results of our algorithm under the uncertainly set.

Theorem 1 (Regret under (s,a)-rectangular uncertainty set)**.** *With learning rate $\beta = \sqrt{\frac{2\log A}{H^2K}}$ and bonus term b_h^k as (4), with probability at least $1-\delta$, the regret incurred by Algorithm 1 over K episodes is bounded by $O\left(H^2S\sqrt{AK\log\left(SAH^2K^{3/2}(1+\rho)/\delta\right)}\right)$.*

Remark 4.1. *When $\rho = 0$, the problem reduces to non-robust reinforcement learning. In such case our regret upper bound is $\tilde{O}\left(H^2S\sqrt{AK}\right)$, which is in the same order of policy optimization algorithms for the non-robust case Shani et al. [30].*

Beyond the (s,a)-rectangular uncertainty set, we also extends to s-rectangular uncertainty set (Definition 2.2).

Algorithm 1 Robust Optimistic Policy Optimization (ROPO)

Input: learning rate β, bonus function b_h^k.
for $k = 1, \ldots, K$ **do**
  Collect a trajectory of samples by executing π_k.
  # Robust Policy Evaluation
  for $h = H, \ldots, 1$ **do**
    for $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
      Solve $\sigma_{\hat{\mathcal{P}}_h}(\hat{V}_{h+1}^{\pi})(s,a)$ according to Equation (2) for (s,a)-rectangular set or Equation (3) for s-rectangular set.
      $\hat{Q}_h^k(s,a) = \min\left\{\hat{r}(s,a) + \sigma_{\hat{\mathcal{P}}_h}(\hat{V}_{h+1}^{\pi})(s,a) + b_h^k(s,a), H\right\}$.
    end for
    for $\forall s \in \mathcal{S}$ **do**
      $\hat{V}_h^k(s) = \left\langle \hat{Q}_h^k(s,\cdot), \pi_h^k(\cdot \mid s)\right\rangle$.
    end for
  end for
  # Policy Improvement
  for $\forall h, s, a \in [H] \times \mathcal{S} \times \mathcal{A}$ **do**
    $\pi_h^{k+1}(a \mid s) = \frac{\pi_h^k \exp(\beta \hat{Q}_h^{\pi}(s,a))}{\sum_{a'} \exp(\beta \hat{Q}_h^{\pi}(s,a'))}$.
  end for
  Update empirical estimate $\hat{r}$, $\hat{P}$ with Equation (1).
end for

Theorem 2 (Regret under s-rectangular uncertainty set)**.** *With learning rate $\beta = \sqrt{\frac{2\log A}{H^2 K}}$ and bonus term b_h^k as (5), with probability at least $1-\delta$, the regret of Algorithm 1 is bounded by $O\left(SA^2H^2\sqrt{K\log(SA^2H^2K^{3/2}(1+\rho)/\delta)}\right)$.*

Remark 4.2. *When $\rho = 0$, the problem reduces to non-robust reinforcement learning. In such case our regret upper bound is $\tilde{O}\left(SA^2H^2\sqrt{K}\right)$. Our result is the first theoretical result for learning a robust policy under s-rectangular uncertainty set, as previous results only learn the robust value function [38].*

We defer the proof of these theorems, along with the experiments results of the proposed algorithm to the Appendix.

5 Conclusion

In this paper, we studied the problem of regret minimization in robust MDP with a rectangular uncertainty set. We proposed a robust variant of optimistic policy optimization, which achieves sublinear regret in all uncertainty sets considered. Our algorithm delicately balances the exploration-exploitation trade-off through a carefully designed bonus term, which quantifies not only the uncertainty due to the limited observations but also the uncertainty of robust MDPs. Our results are the first regret upper bounds in robust MDPs as well as the first non-asymptotic results in robust MDPs without access to a generative model.

Acknowledgement

Jing Dong and Baoxiang Wang are partially supported by National Natural Science Foundation of China (62106213, 72150002) and Shenzhen Science and Technology Program (RCBS20210609104356063, JCYJ20210324120011032). Jingzhao Zhang is supported by Tsinghua University Initiative Scientific Research Program.

References

[1] Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, pages 10–4, 2019.

[2] Kishan Panaganti Badrinath and Dileep Kalathil. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *International Conference on Machine Learning*, 2021.

[3] Peter Bartlett. Theoretical statistics. lecture 12, 2013.

[4] Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, 2020.

[5] Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Minimax regret for stochastic shortest path with adversarial costs and known transition. In *Conference on Learning Theory*, 2021.

[6] Yifang Chen, Simon Du, and Kevin Jamieson. Improved corruption robust algorithms for episodic reinforcement learning. In *International Conference on Machine Learning*, 2021.

[7] Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *International Conference on Machine Learning*, 2019.

[8] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 2017.

[9] Omar Darwiche Domingues, Yannis Flet-Berliac, Edouard Leurent, Pierre Ménard, Xuedong Shang, and Michal Valko. rlberry - A Reinforcement Learning Library for Research and Education, 10 2021. URL `https://github.com/rlberry-py/rlberry`.

[10] Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Experts in a Markov decision process. *Advances in Neural Information Processing Systems*, 2004.

[11] Jesse Farebrother, Marlos C Machado, and Michael Bowling. Generalization and regularization in DQN. *arXiv preprint arXiv:1810.00123*, 2018.

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

[13] Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial Markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, 2020.

[14] Tiancheng Jin and Haipeng Luo. Simultaneously learning stochastic and adversarial episodic MDPs with known transition. *Advances in Neural Information Processing Systems*, 2020.

[15] Nathan Kallus, Xiaojie Mao, Kaiwen Wang, and Zhengyuan Zhou. Doubly robust distributionally robust off-policy evaluation and learning. *International Conference on Machine Learning*, 2022.

[16] Zijian Liu, Qinxun Bai, Jose Blanchet, Perry Dong, Wei Xu, Zhengqing Zhou, and Zhengyuan Zhou. Distributionally robust q-learning. In *International Conference on Machine Learning*, pages 13623–13643. PMLR, 2022.

[17] Thodoris Lykouris, Max Simchowitz, Alex Slivkins, and Wen Sun. Corruption-robust exploration in episodic reinforcement learning. In *Conference on Learning Theory*, 2021.

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

[19] Shie Mannor, Ofir Mebel, and Huan Xu. Lightning does not strike twice: robust MDPs with coupled uncertainty. In *International Conference on Machine Learning*, 2012.

[20] Gergely Neu, Andras Antos, András György, and Csaba Szepesvári. Online Markov decision processes under bandit feedback. *Advances in Neural Information Processing Systems*, 2010.

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

[22] Charles Packer, Katelyn Gao, Jernej Kos, Philipp Krähenbühl, Vladlen Koltun, and Dawn Song. Assessing generalization in deep reinforcement learning. *arXiv preprint arXiv:1810.12282*, 2018.

[23] Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, 2022.

[24] Zhengling Qi and Peng Liao. Robust batch policy learning in Markov decision processes. *arXiv preprint arXiv:2011.04185*, 2020.

[25] Roberta Raileanu and Rob Fergus. Decoupling value and policy for generalization in reinforcement learning. In *International Conference on Machine Learning*, 2021.

[26] Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial Markov decision processes. In *International Conference on Machine Learning*, 2019.

[27] Jay K Satia and Roy E Lave Jr. Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740, 1973.

[28] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, 2015.

[29] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[30] Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, 2020.

[31] Xingyou Song, Yiding Jiang, Stephen Tu, Yilun Du, and Behnam Neyshabur. Observational overfitting in reinforcement learning. In *International Conference on Learning Representations*, 2019.

[32] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

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

[34] Yue Wang and Shaofeng Zou. Policy gradient method for robust reinforcement learning. *International Conference on Machine Learning*, 2022.

[35] Chelsea C White III and Hany K Eldeib. Markov decision processes with imprecise transition probabilities. *Operations Research*, 42(4):739–749, 1994.

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

[37] Tianhao Wu, Yunchang Yang, Han Zhong, Liwei Wang, Simon Du, and Jiantao Jiao. Nearly optimal policy optimization with stable at any time guarantee. In *International Conference on Machine Learning*, 2022.

[38] Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Towards theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics. *arXiv preprint arXiv:2105.03863*, 2021.

[39] Xuezhou Zhang, Yiding Chen, Xiaojin Zhu, and Wen Sun. Corruption-robust offline reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, 2022.

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

A Importance of robustness

With the robust MDP, one of the most naive methods is to directly train a policy with the nominal transition model. However, the following proposition shows an optimal policy under the nominal policy can be arbitrarily bad in the worst-case transition (even worse than a random policy).

Claim A.1 (Suboptimality of non-robust optimal policy)**.** *There exists a robust MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, r, H \rangle$ with uncertainty set $\mathcal{P}$ of uncertainty radius ρ, such that the non-robust optimal policy is $\Omega(1)$-suboptimal to the uniformly random policy.*

The proof of Proposition A.1 is deferred to Appendix G. With the above-stated result, it implies the policy obtained with non-robust RL algorithms, can have arbitrarily bad performance when the dynamic mismatch from the nominal transition. This thus motivate our robust optimistic policy optimization 1 to avoid this undesired result.

B Related works

RL with robust MDP Different from conventional MDPs, robust MDPs allow the transition kernel to take values from an uncertainty set. The objective in robust MDPs is to learn an optimal robust policy that maximizes the worst-case value function. When the exact uncertainty set is known, this can be solved through dynamic programming methods [12, 21, 19]. Yet knowing the exact uncertainty set is a rather stringent requirement for most real applications. If one has access to a generative model, several model-based reinforcement learning methods are proven to be statistically efficient. With the different characterization of the uncertainty set, these methods can enjoy a sample complexity of $O(1/\epsilon^2)$ for an ϵ-optimal robust value function [23, 38]. Similar results can also be achieved if an offline dataset is present, for which previous works Qi and Liao [24], Zhou et al. [40], Kallus et al. [15], Ma et al. [18] show the $O(1/\epsilon^2)$ sample complexity for an ϵ-optimal policy. In addition, Liu et al. [16] proposed distributionally robust policy Q-learning, which solves for the asymptotically optimal Q-function.

In the case of online RL, the only results available are asymptotic. In the case of discounted MDPs, Wang and Zou [33], Badrinath and Kalathil [2] study the policy gradient method and show an $O(\epsilon^{-3})$ convergence rate for an alternative learning objective (a smoothed variant), which could be equivalent to the original policy gradient objective in an asymptotic regime. These results in sample complexity and asymptotic regimes in general cannot imply sublinear regret in robust MDPs [8].

RL with adversarial MDP Another line of works characterizes the uncertainty of the environment through the adversarial MDP formulation, where the environmental parameters can be adversarially chosen without restrictions. This problem is proved to be NP-hard to obtain a low regret [10]. Several works study the variant where the adversarial could only modify the reward function, while the transition dynamics of the MDP remain unchanged. In this case, it is possible to obtain policy-based algorithms that are efficient with a sublinear regret [26, 14, 13, 30, 4]. On a separate vein, it investigates the setting where the transition is only allowed to be adversarially chosen for C out of the K total episodes. A regret of $O(C^2 + \sqrt{K})$ are established thereafter [17, 6, 39].

Non-robust policy optimization The problem of policy optimization has been extensively investigated under non-robust MDPs [20, 4, 30, 37, 5]. The proposed methods are proved to achieve sublinear regret. The methods are also closely related to empirically successful policy optimization algorithms in RL, such as PPO Schulman et al. [29] and TRPO Schulman et al. [28].

C Experiments

To validate our theoretical findings, we conduct a preliminary empirical analysis of our purposed robust policy optimization algorithm.

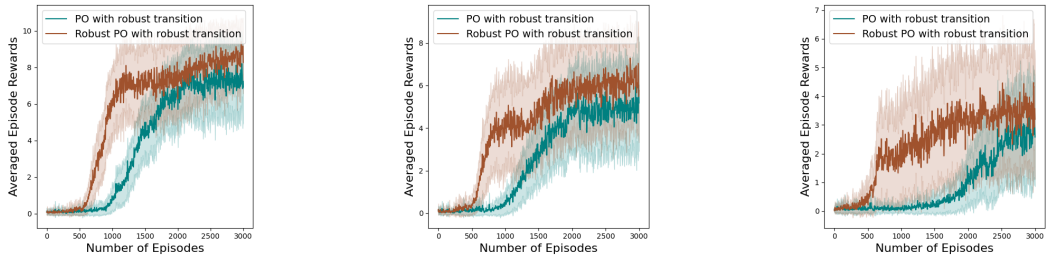
Environment We conduct the experiments with the Gridworld environment, which is an early example of reinforcement learning from [32]. The environment is two-dimensional and is in a cell-like environment. Specifically, the environment is a 5×5 grid, where the agent starts from the upper left cell. The cells consist of three types, road (labeled with o), wall (labeled with x), or reward state (labeled with $+$). The agent can safely walk through the road cell but not the wall cell. Once the agent steps on the reward cell, it will receive a reward of 1, and it will receive no rewards otherwise. The goal of the agents is to collect as many rewards as possible within the allowed time. The agent has four types of actions at each step, up, down, left, and right. After taking the action, the agent has a success probability of p to move according to the desired direction, and with the remaining probability of moving to other directions.

Start	o	o	o	o
o	x	o	o	o
o	o	x	o	o
o	o	o	x	o
o	o	o	o	+

Figure 1: Example of the Gridworld environment.

Experiment configurations To simulate the robust MDP, we create a nominal transition dynamic with success probability $p = 0.9$. The learning agent will interact with this nominal transition during training time and interact with a perturbed transition dynamic during evaluation.Under (s, a)-rectangular set, the transitions are perturbed against the direction is agent is directing with a constraint of ρ. Under s-rectangular set, the transitions are perturbed against the direction of the goal state. Figure 1 shows an example of our environment, where the perturbation caused some of the optimal policies under nominal transition to be sub-optimal under robust transitions. We denote the perturbed transition as robust transitions in our results. We implement our proposed robust policy optimization algorithm along with the non-robust variant of it [30]. The inner minimization of our Algorithm 1 is computed through its dual formulation for efficiency. Our algorithm is implemented with the rLberry framework [9].

Results We present results with $\rho = 0.1, 0.2, 0.3$ under (s, a)-rectangular set here in Figure 4. The results with s-rectangular sets are included in the appendix. We present the averaged cumulative rewards during evaluation. Regardless of the level of uncertainty, we observe that the robust variant of the policy optimization algorithm is more robust to dynamic changes as it is able to obtain a higher level of rewards than its non-robust variant.



(a) $\rho = 0.1$ **(b)** $\rho = 0.2$ **(c)** $\rho = 0.3$

Figure 2: Cumulative rewards obtained by robust and non-robust policy optimization on robust transition with different level of uncertainty $\rho = 0.1, 0.2, 0.3$ under ℓ_1 distance, (s, a)-rectangular set.

D Proofs of Theorem 1

D.1 Good events

We first define the following good events, in which case we estimate the reward function and the nominal transition functions fairly accurately.

$$\mathcal{G}_k^r = \left\{ \forall s,a,h : \left| r_h(s,a) - \hat{r}_h^k(s,a) \right| \le \sqrt{\frac{2\ln(2SAH^2K/\delta')}{N_h^k(s,a)}} \right\},$$

$$\mathcal{G}_k^p = \left\{ \forall s,a,h : \sigma_{\mathcal{P}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s,a) - \sigma_{\hat{\mathcal{P}}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s,a) \le C_h^k(s,a) \right\},$$

where $C_h^k(s,a) = H\sqrt{\frac{4S\log(3SAH^2K^{3/2}(4+\rho)/\delta')}{N_h^k(s,a)}} + \frac{1}{\sqrt{K}}$.

When the two good events happens at the same time, we say the algorithm in inside the good event $\mathcal{G} = \left(\bigcap_{k=1}^K \mathcal{G}_k^r\right) \bigcap \left(\bigcap_{k=1}^K \mathcal{G}_k^p\right)$. The following lemma shows that $\mathcal{G}$ happens with high probability by setting δ' properly.

Lemma 1 (Good event)**.** *Let $\delta = 2\delta'$, then the good event happens with high probability, i.e. $\mathbb{P}[\mathcal{G}] \ge 1-\delta$.*

Proof. By Hoeffding's inequality and an union bound on all s, a, all possible values of $N_k(s,a)$ and k, we have $\mathbb{P}\left[\bigcap_{k=1}^K \mathcal{G}_k^r\right] \ge 1-\delta'$. By Lemma 4, we have $\mathbb{P}\left[\bigcap_{k=1}^K \mathcal{G}_k^p\right] \ge 1-\delta'$ Then set $\delta = 2\delta'$ and we have the desired result. □

D.2 Design of the bonus function

In the case of (s,a)-rectangular uncertainty set, we use the following bonus function $b_h^k(s,a)$ to encourage exploration.

$$b_h^k(s,a) = \sqrt{\frac{2\log(3SAH^2K/\delta)}{N_h^k(s,a)}} + H\sqrt{\frac{4S\log(3SAH^2K^{3/2}(4+\rho)/\delta)}{N_h^k(s,a)}} + \frac{1}{\sqrt{K}}. \tag{4}$$

D.3 Regret Analysis

Armed with the defined good event, we are now ready to present the anlysis of Theorem 1, which establishes the regret of the Algorithm under (s,a)-uncertainty set.

Theorem 1 (Regret under (s,a)-rectangular uncertainty set)**.** *With learning rate $\beta = \sqrt{\frac{2\log A}{H^2K}}$ and bonus term b_h^k as (4), with probability at least $1-\delta$, the regret incurred by Algorithm 1 over K episodes is bounded by $O\left(H^2S\sqrt{AK\log\left(SAH^2K^{3/2}(1+\rho)/\delta\right)}\right)$.*

Proof. We start with decomposing the regret as follows,

$$\begin{aligned}\text{Regret}(K) &= \sum_{k=1}^K V_1^*(s) - V_1^{\pi_k}(s) \\ &= \sum_{k=1}^K \left(V_1^*(s) - \hat{V}_1^{\pi_k}(s)\right) + \left(\hat{V}_1^{\pi_k}(s) - V_1^{\pi_k}(s)\right).\end{aligned}$$

By Lemma 2 and Lemma 3, with probability at least $1-\delta$, we have

$$\begin{aligned}\text{Regret}(K) &= O\left(H^2\sqrt{K\log A}\right) + O\left(H^2S\sqrt{AK\log\left(SAH^2K^{3/2}(1+\rho)/\delta\right)}\right) \\ &= O\left(H^2S\sqrt{AK\log\left(SAH^2K^{3/2}(1+\rho)/\delta\right)}\right).\end{aligned}$$

□

Lemma 2. *With probability at least* $1-\delta$*, we have*

$$\sum_{k=1}^{K} V_1^*(s) - \hat{V}_1^{\pi_k}(s) = O\left(H^2\sqrt{K\log A}\right).$$

Proof. For any $h \in [1, H]$, we have

$$\begin{aligned}
&V_h^*(s) - \hat{V}_h^{\pi_k}(s) \\
&= \langle Q_h^*(s,\cdot), \pi_*(\cdot \mid s)\rangle - \langle \hat{Q}_h^{\pi_k}(s,\cdot), \pi_k(\cdot \mid s)\rangle \\
&= \langle Q_h^*(s,\cdot) - \hat{Q}_h^{\pi_k}(s,\cdot), \pi_*(\cdot \mid s)\rangle + \langle \hat{Q}_h^{\pi_k}(s,\cdot), \pi_*(\cdot \mid s) - \pi_k(\cdot \mid s)\rangle \\
&= \mathbb{E}_{\pi_*}\left[(r_h(s,a) - \hat{r}_h^k(s,a)) + (\sigma_{\mathcal{P}_h(s,a)}(V_{h+1}^*)(s,a) - \sigma_{\hat{\mathcal{P}}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s,a)) - b_h^k(s,a)\right] \\
&\quad + \langle \hat{Q}_h^{\pi_k}(s,\cdot), \pi_*(\cdot \mid s) - \pi_k(\cdot \mid s)\rangle \\
&= \mathbb{E}_{\pi_*}\left[(r_h(s,a) - \hat{r}_h^k(s,a)) + (\sigma_{\mathcal{P}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s,a) - \sigma_{\hat{\mathcal{P}}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s,a)) - b_h^k(s,a)\right] \\
&\quad + \mathbb{E}_{\pi_*}\left[\sigma_{\mathcal{P}_h(s,a)}(V_{h+1}^*)(s,a) - \sigma_{\mathcal{P}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s,a)\right] + \langle \hat{Q}_h^{\pi_k}(s,\cdot), \pi_*(\cdot \mid s) - \pi_k(\cdot \mid s)\rangle,
\end{aligned}$$

where the third equality is by the update rule of our algorithm and the robust bellman equation. By the design of our bonus function, conditioned on the good event, we have

$$(r_h(s,a) - \hat{r}_h^k(s,a)) + (\sigma_{\mathcal{P}_h(s,a)}(V_{h+1}^*)(s,a) - \sigma_{\hat{\mathcal{P}}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s,a)) - b_h^k(s,a) \le 0.$$

Let $q_h(\cdot \mid s,a) = \arg\min_{P_h \in \mathcal{P}_h} P_h(\cdot \mid s,a)\hat{V}_{h+1}^{\pi_k}$, then we have

$$\begin{aligned}
&\sigma_{\mathcal{P}_h(s,a)}(V_{h+1}^*)(s,a) - \sigma_{\mathcal{P}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s,a) \\
&= \min_{P_h \in \mathcal{P}_h} P_h(\cdot \mid s,a)V_{h+1}^* - \min_{P_h \in \mathcal{P}_h} P_h(\cdot \mid s,a)\hat{V}_{h+1}^{\pi_k} \\
&= \min_{P_h \in \mathcal{P}_h} P_h(\cdot \mid s,a)V_{h+1}^* - q_h(\cdot \mid s,a)\hat{V}_{h+1}^{\pi_k} \\
&\le q_h(\cdot \mid s,a)(V_{h+1}^* - \hat{V}_{h+1}^{\pi_k}) \\
&\le \max_{P_h \in \mathcal{P}_h} P_h(\cdot \mid s,a)(V_{h+1}^* - \hat{V}_{h+1}^{\pi_k}).
\end{aligned}$$

Let $p_h(\cdot \mid s,a) = \arg\max_{P_h \in \mathcal{P}_h} P_h(\cdot \mid s,a)(V_{h+1}^*)(s,a)$, Then we have the following relation hold conditioned on the good event:

$$\begin{aligned}
&V_h^*(s) - \hat{V}_h^{\pi_k}(s) \\
&\le \mathbb{E}_{\pi_*}\left[\sup_{P_h \in \mathcal{P}_h} P_h(\cdot \mid s,a)(V_{h+1}^* - \hat{V}_{h+1}^{\pi_k})\right] + \langle \hat{Q}_h^{\pi_k}(s,\cdot), \pi_*(\cdot \mid s) - \pi_k(\cdot \mid s)\rangle \\
&= \mathbb{E}_{\pi_*, p_h}\left[V_{h+1}^*(s) - \hat{V}_{h+1}^{\pi_k}(s)\right] + \langle \hat{Q}_h^{\pi_k}(s,\cdot), \pi_*(\cdot \mid s) - \pi_k(\cdot \mid s)\rangle.
\end{aligned}$$

Then, by applying above relation recursively and with the fact that for any policy π and state s, $V_{H+1}^*(s) = \hat{V}_{H+1}^{\pi_k}(s) = 0$, we have

$$V_1^*(s) - \hat{V}_1^{\pi_k}(s) \le \sum_{h=1}^{H} \mathbb{E}_{\pi_*, \{q_t\}_{t=1}^{h-1}}\left[\langle \hat{Q}_h^{\pi_k}(s,\cdot), \pi_*(\cdot \mid s) - \pi_k(\cdot \mid s)\rangle\right].$$

Summing over k, we get

$$\sum_{k=1}^{K} V_1^*(s) - \hat{V}_1^{\pi_k}(s) \le \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{\pi_*, \{q_t\}_{t=1}^{h-1}}\left[\langle \hat{Q}_h^{\pi_k}(s,\cdot), \pi_*(\cdot \mid s) - \pi_k(\cdot \mid s)\rangle\right]$$

$$= \sum_{h=1}^{H} \mathbb{E}_{\pi_*, \{q_t\}_{t=1}^{h-1}} \left[\sum_{k=1}^{K} \langle \hat{Q}_h^{\pi_k}(s, \cdot), \pi_*(\cdot \mid s) - \pi_k(\cdot \mid s) \rangle \right] .$$

By standard results for online mirror descent (Lemma 13), we have

$$\sum_{k=1}^{K} \langle \hat{Q}_h^{\pi_k}(s, \cdot), \pi_*(\cdot \mid s) - \pi_k(\cdot \mid s) \rangle \leq \frac{\log(A)}{\beta} + \frac{\beta}{2} \sum_{k=1}^{K} \sum_{a \in \mathcal{A}} \pi_h^*(a \mid s) (\hat{Q}_h^{\pi_k}(s, a))^2 .$$

By the update rule of Algorithm 1, we have $0 \leq \hat{Q}_h^{\pi_k}(s, a) \leq H$, for all h, k. Then take $\beta = \sqrt{\frac{2 \log A}{H^2 K}}$,

$$\sum_{k=1}^{K} \langle \hat{Q}_h^{\pi_k}(s, \cdot), \pi_*(\cdot \mid s) - \pi_k(\cdot \mid s) \rangle \leq \sqrt{2 H^2 K \log A} .$$

Finally, we have

$$\sum_{k=1}^{K} V_1^*(s) - \hat{V}_1^{\pi_k}(s) \leq H \sqrt{2 H^2 K \log A} = O\left(H^2 \sqrt{K \log A} \right) .$$

□

Lemma 3. *With probability at least* $1 - \delta$*, we have*

$$\sum_{k=1}^{K} (\hat{V}_1^{\pi_k} - V_1^{\pi_k})(s) = O\left(H^2 S \sqrt{AK \log\left(SAH^2 K^{3/2} (1 + \rho) / \delta \right)} \right) .$$

Proof. By the algorithm's update rule and the robust bellman equation, we have

$$\begin{aligned}
(\hat{V}_h^{\pi_k} - V_h^{\pi_k})(s) &= \langle \hat{Q}_h^{\pi_k}(s, \cdot) - Q_h^{\pi_k}(s, \cdot), \pi_k(\cdot \mid s) \rangle \\
&= \left\langle \hat{r}_h^k(s, \cdot) - r_h^k(s, \cdot) + (\sigma_{\hat{\mathcal{P}}_{(s,\cdot)}}(\hat{V}_{h+1}^{\pi_k})(s, \cdot) - \sigma_{\mathcal{P}_{(s,\cdot)}}(V_{h+1}^{\pi_k})(s, \cdot)) + b_h^k(s, \cdot), \pi_k(\cdot \mid s) \right\rangle \\
&= \mathbb{E}_{\pi_k} \left[\hat{r}_h^k(s, a) - r_h^k(s, a) + (\sigma_{\hat{\mathcal{P}}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s, a) - \sigma_{\mathcal{P}_h(s,a)}(V_{h+1}^{\pi_k})(s, a)) + b_h^k(s, a) \right] .
\end{aligned}$$

By adding and subtracting a term $\sigma_{\mathcal{P}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s, a)$, we have

$$\begin{aligned}
&\sigma_{\hat{\mathcal{P}}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s, a) - \sigma_{\mathcal{P}_h(s,a)}(V_{h+1}^{\pi_k})(s, a) \\
&= \sigma_{\hat{\mathcal{P}}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s, a) - \sigma_{\mathcal{P}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s, a) + \sigma_{\mathcal{P}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s, a) - \sigma_{\mathcal{P}_h(s,a)}(V_{h+1}^{\pi_k})(s, a) \\
&\leq \sigma_{\hat{\mathcal{P}}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s, a) - \sigma_{\mathcal{P}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s, a) + \max_{P_h \in \mathcal{P}_h} P_h(\cdot \mid s, a)(\hat{V}_{h+1}^{\pi_k} - V_{h+1}^{\pi_k}) .
\end{aligned}$$

Let $p_h(\cdot \mid s, a) = \arg\max_{P_h \in \mathcal{P}_h} P_h(\cdot \mid s, a)(\hat{V}_{h+1}^{\pi_k} - V_{h+1}^{\pi_k})$, we have

$$\begin{aligned}
&(\hat{V}_h^{\pi_k} - V_h^{\pi_k})(s) \\
&\leq \mathbb{E}_{\pi_k} \left[\hat{r}_h^k(s, a) - r_h^k(s, a) + \sigma_{\hat{\mathcal{P}}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s, a) - \sigma_{\mathcal{P}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s, a) + p_h(\cdot \mid s, a)(\hat{V}_{h+1}^{\pi_k} - V_{h+1}^{\pi_k}) + b_h^k(s, a) \right] \\
&= \mathbb{E}_{\pi_k, p_h} \left[\hat{r}_h^k(s, a) - r_h^k(s, a) + \sigma_{\hat{\mathcal{P}}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s, a) - \sigma_{\mathcal{P}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s, a) + \hat{V}_{h+1}^{\pi_k}(s) - V_{h+1}^{\pi_k}(s) + b_h^k(s, a) \right]
\end{aligned}$$

By applying the above relation recursively and with the fact that for any policy π and state s, $V_{H+1}^{\pi_k}(s) = \hat{V}_{H+1}^{\pi_k}(s) = 0$, we have

$$(\hat{V}_1^{\pi_k} - V_1^{\pi_k})(s) \leq \sum_{h=1}^{H} \mathbb{E}_{\pi_k, \{p_t\}_{t=1}^{h}} \left[\hat{r}_h^k(s, a) - r_h^k(s, a) + \sigma_{\hat{\mathcal{P}}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s, a) - \sigma_{\mathcal{P}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s, a) + b_h^k(s, a) \right] .$$

Conditioned on the good even and by the design of our bonus function, we have

$$\hat{r}_h^k(s,a) - r_h^k(s,a) + \sigma_{\hat{\mathcal{P}}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s,a) - \sigma_{\mathcal{P}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s,a) \le b_h^k(s,a)\,.$$

Then, with probability at least $1-\delta$, we have

$$\begin{aligned}\sum_{k=1}^{K}(\hat{V}_1^{\pi_k} - V_1^{\pi_k})(s) &\le \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi_k,\{p_t\}_{t=1}^h}\left[2b_h^k(s,a)\right] \\ &\le H\sqrt{K} + O\left(H\sqrt{S\log(SAH^2K^{3/2}(4+\rho)/\delta)}\right)\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi_k,\{p_t\}_{t=1}^h}\left[\sqrt{\frac{1}{N_h^k(s,a)}}\right]\,.\end{aligned}$$

By Lemma 12, we have the bound of the visitation counts:

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{1}{N_h^k(s,a)}} \le 2H\sqrt{SAK}\,.$$

Combining everything, with probability at least $1-\delta$

$$\sum_{k=1}^{K}(\hat{V}_1^{\pi_k} - V_1^{\pi_k})(s) = O\left(H^2S\sqrt{AK\log\left(SAH^2K^{3/2}(1+\rho)/\delta\right)}\right)\,.$$

□

Lemma 4. *For any h,k,s,a, the following inequality holds with probability at least $1-\delta'$,*

$$\sigma_{\mathcal{P}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s,a) - \sigma_{\hat{\mathcal{P}}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s,a) \le H\sqrt{\frac{4S\log(3SAH^3K^{3/2}(4+\rho)/\delta')}{N_h^k(s,a)}} + \frac{1}{H\sqrt{K}}\,.$$

Proof. By the definition of $\sigma_{\mathcal{P}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s,a) = \min_{P_h\in\mathcal{P}_h}\sum_{s'}P_h(s'\mid s,a)\hat{V}_{h+1}^{\pi_k}(s')$, we have the following optimization problem:

$$\begin{aligned}&\min_{P_h}\ \sum_{s'}P_h(s'\mid s,a)\hat{V}_{h+1}^{\pi_k}(s') \\ &\text{s.t.}\quad \begin{cases}\sum_{s'}|P_h(s'\mid s,a) - P_h^o(s'\mid s,a)| \le \rho\,, \\ \sum_{s'}P_h(s'\mid s,a) = 1\,, \\ P_h^o(\cdot\mid s,a) > 0, P_h(\cdot\mid s,a)\ge 0\,.\end{cases}\end{aligned}$$

Define $\tilde{P}_h(s'\mid s,a) = \frac{P_h(s'|s,a)}{P_h^o(s'|s,a)}$, we can rewrite the above optimization problem as

$$\begin{aligned}&\min_{\tilde{P}_h}\sum_{s'}\tilde{P}_h(s'\mid s,a)P_h^o(s'\mid s,a)\hat{V}_{h+1}^{\pi_k}(s') \\ &\text{s.t.}\quad \begin{cases}\sum_{s'}|\tilde{P}_h(s'\mid s,a) - 1|P_h^o(s'\mid s,a) \le \rho\,, \\ \sum_{s'}\tilde{P}_h(s'\mid s,a)P_h^o(s'\mid s,a) = 1\,, \\ \tilde{P}_h(s'\mid s,a)\ge 0 \quad \forall s'\in\mathcal{S}\,.\end{cases}\end{aligned}$$

Using the Lagrangian multiplier method, we have the following Lagrangian $L(\tilde{P}_h,\eta,\lambda)$ with Lagrangian multiplier $\eta\in\mathbb{R},\lambda\ge 0$,

$$\begin{aligned}L(\tilde{P}_h,\eta,\lambda)(s,a) = &\sum_{s'}\tilde{P}_h(s'\mid s,a)P_h^o(s'\mid s,a)\hat{V}_{h+1}^{\pi_k}(s') + \lambda\left(\sum_{s'}|\tilde{P}_h(s'\mid s,a)-1|P_h^o(s'\mid s,a) - \rho\right) \\ &- \eta\left(\sum_{s'}\tilde{P}_h(s'\mid s,a)P_h^o(s'\mid s,a) - 1\right)\end{aligned}$$

$$= \eta - \lambda\rho - \lambda \sum_{s'} P_h^o(s' \mid s,a) \left(\frac{\eta}{\lambda} \tilde{P}_h(s' \mid s,a) - |\tilde{P}_h(s' \mid s,a) - 1| - \frac{\tilde{P}_h(s' \mid s,a) \hat{V}_{h+1}^{\pi_k}(s')}{\lambda} \right)$$

$$= \eta - \lambda\rho - \lambda \sum_{s'} P_h^o(s' \mid s,a) \left(\frac{\eta - \hat{V}_{h+1}^{\pi_k}(s')}{\lambda} \tilde{P}_h(s' \mid s,a) - |\tilde{P}_h(s' \mid s,a) - 1| \right) .$$

We define $f(x) = |x-1|$ and the convex conjugate is $f^*(y) = \max_x \langle x, y \rangle - f(x)$. Let x be $\tilde{P}_h$ and by using f^*, we can optimize over $\tilde{P}_h$ and rewrite the Lagrangian as

$$L(\eta, \lambda)(s,a) = \min_{\tilde{P}_h} L(\tilde{P}_h, \eta, \lambda)(s,a) = \eta - \lambda\rho - \lambda \sum_{s'} P_h^o(s' \mid s,a) f^* \left(\frac{\eta - \hat{V}_{h+1}^{\pi_k}(s')}{\lambda} \right) .$$

Notice that conditioned on $x \geq 0$, $f(x) = |x-1|$'s convex conjugate has the following closed form:

$$f^*(y) = \max_x \langle x, y \rangle - f(x) = \begin{cases} -1 & y \leq -1\,, \\ y & y \in [-1, 1]\,, \\ +\infty & y > 1\,. \end{cases}$$

Let $\tilde{\eta} = \eta + \lambda$, then using the closed form of $f^*(y)$, the equality $\max\{a, b\} = (a-b)_+ + b$ and condition on $\frac{\eta - \hat{V}_{h+1}^{\pi_k}(s')}{\lambda} \leq 1$, we can rewrite the optimization problem as

$$\begin{aligned} L(\tilde{\eta}, \lambda)(s,a) &= \eta - \lambda\rho - \lambda \sum_{s'} P_h^o(s' \mid s,a) f^* \left(\frac{\eta - \hat{V}_{h+1}^{\pi_k}(s')}{\lambda} \right) \\ &= \tilde{\eta} - \lambda - \lambda\rho - \lambda \sum_{s'} P_h^o(s' \mid s,a) \max \left\{ \frac{\eta - \hat{V}_{h+1}^{\pi_k}(s')}{\lambda}, -1 \right\} \\ &= \tilde{\eta} - \lambda - \lambda\rho - \lambda \sum_{s'} P_h^o(s' \mid s,a) \left(\left(\frac{\eta - \hat{V}_{h+1}^{\pi_k}(s')}{\lambda} - (-1) \right)_+ + (-1) \right) \\ &= \tilde{\eta} - \lambda - \lambda\rho - \sum_{s'} P_h^o(s' \mid s,a) (\tilde{\eta} - \hat{V}_{h+1}^{\pi_k}(s'))_+ + \lambda \\ &= \tilde{\eta} - \lambda\rho - \sum_{s'} P_h^o(s' \mid s,a) (\tilde{\eta} - \hat{V}_{h+1}^{\pi_k}(s'))_+ \,. \end{aligned}$$

with the constraint of λ being

$$\lambda \geq 0, \quad \tilde{\eta} - \min_s \hat{V}_{h+1}^{\pi_k}(s) \leq 2\lambda.$$

Then we discuss the constraint of $\tilde{\eta} = \eta + \lambda$ and show that $\tilde{\eta} \in R$. We discuss this by cases.

For any $x \leq \min_s \hat{V}_{h+1}^{\pi_k}(s)$, taking $\eta = x$, $\lambda = 0$, then we have $\tilde{\eta} = x$.

For any $x > \min_s \hat{V}_{h+1}^{\pi_k}(s)$, taking $\eta = \frac{x + \min_s \hat{V}_{h+1}^{\pi_k}(s)}{2}$, $\lambda = \frac{x - \min_s \hat{V}_{h+1}^{\pi_k}(s)}{2}$, then we have $\tilde{\eta} = x$.

Then we have $\tilde{\eta} \in R$. Fixing any $\tilde{\eta}$, from the definition of L, we need to choose $\lambda = \frac{(\tilde{\eta} - \min_s \hat{V}_{h+1}^{\pi_k}(s))_+}{2}$ to achieve the maximum of L. Then by directly optimizing it over λ, we can reduce the problem to

$$L(\tilde{\eta})(s,a) = \tilde{\eta} - \frac{(\tilde{\eta} - \min_s \hat{V}_{h+1}^{\pi_k}(s))_+}{2} \rho - \sum_{s'} P_h^o(s' \mid s,a) (\tilde{\eta} - \hat{V}_{h+1}^{\pi_k}(s'))_+ \,.$$

with the constraint $\tilde{\eta} \in R$.

Define the function g as

$$g(\tilde{\eta}, P_h^o) = -L(\tilde{\eta})(s,a) = \sum_{s'} P_h^o(s' \mid s,a) \left(\tilde{\eta} - \hat{V}_{h+1}^{\pi_k}(s') \right)_+ - \tilde{\eta} + \frac{(\tilde{\eta} - \min_s \hat{V}_{h+1}^{\pi_k}(s))_+}{2} \rho \,.$$

Then we investigate the optimum of g. First notice that $g(0) = 0$, when $\tilde{\eta} \leq 0$, $g(\tilde{\eta}, P_h^o) = -\tilde{\eta} \geq 0$. On the other hand, when $\tilde{\eta} \geq H$,

$$\begin{aligned} g(\tilde{\eta}, P_h^o) &= \sum_{s'} P_h^o(s' \mid s,a)(\tilde{\eta} - \hat{V}_{h+1}^{\pi_k}(s')) - \tilde{\eta} + \frac{(\tilde{\eta} - \min_s \hat{V}_{h+1}^{\pi_k}(s))}{2}\rho \\ &= -\sum_{s'} P_h^o(s' \mid s,a)\hat{V}_{h+1}^{\pi_k}(s') + \frac{(\tilde{\eta} - \min_s \hat{V}_{h+1}^{\pi_k}(s))}{2}\rho\,. \end{aligned}$$

Note that now g is directly proportional to $\tilde{\eta}$, therefore g achieves the minimum within the range of $\tilde{\eta} \in [0, H]$. We remark that the same form is also used for analyzing robust policy evaluation (Lemma B.1 [38]).

With this, we can rewrite

$$\begin{aligned} \sigma_{\hat{\mathcal{P}}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s,a) - \sigma_{\mathcal{P}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s,a) &= -\min_{\eta_1 \in [0,H]} g(\eta_1, \hat{P}_h^{o,k}) + \min_{\eta_2 \in [0,H]} g(\eta_2, P_h^o) \\ &\leq \max_{\eta \in [0,H]} |g\left(\eta, \hat{P}_h^{o,k}\right) - g(\eta, P_h^o)|\,. \end{aligned}$$

To upper bound $\sigma_{\hat{\mathcal{P}}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s,a) - \sigma_{\mathcal{P}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s,a)$, we first upper bound $|g\left(\eta, \hat{P}_h^{o,k}\right) - g(\eta, P_h^o)|$.

$$\begin{aligned} |g\left(\eta, \hat{P}_h^{o,k}\right) - g(\eta, P_h^o)| &= \left| \sum_{s'} \hat{P}_h^{o,k}(s' \mid s,a)\left(\eta - \hat{V}_{h+1}^{\pi_k}(s')\right)_+ - \sum_{s'} P_h^o(s' \mid s,a)\left(\eta - \hat{V}_{h+1}^{\pi_k}(s')\right)_+ \right| \\ &\leq \left\| \hat{P}_h^{o,k}(\cdot \mid s,a) - P_h^o(\cdot \mid s,a) \right\|_1 \max_{s \in \mathcal{S}} |\eta - \hat{V}_{h+1}^{\pi_k}(s)|_\infty \\ &\leq H \left\| \hat{P}_h^{o,k}(\cdot \mid s,a) - P_h^o(\cdot \mid s,a) \right\|_1, \end{aligned}$$

where the first inequality is by Cauchy-Schwarz inequality, the second inequality follows from $\eta \in [0, H]$.

By Hoeffding's inequality and an union bound over all s, a, the following inequality holds with probability at least $1 - \delta'$:

$$\left\| \hat{P}_h^{o,k}(\cdot \mid s,a) - P_h^o(\cdot \mid s,a) \right\|_1 \leq \sqrt{\frac{4S\log(3SAH^2K/\delta')}{N_h^k(s,a)}}\,.$$

To upper bound the error with maximum over η, we first create an ϵ-net $N_\epsilon(\eta)$ with g over $\eta \in [0, H]$ such that

$$\max_{\eta \in [0,H]} |g\left(\eta, \hat{P}_h^{o,k}\right) - g(\eta, P_h^o)| \leq \max_{\eta \in N_\epsilon(\eta)} |g\left(\eta, \hat{P}_h^{o,k}\right) - g(\eta, P_h^o)| + 2\epsilon\,.$$

By taking an union bound over $N_\epsilon(\eta)$, we have

$$\max_{\eta \in [0,H]} |g\left(\eta, \hat{P}_h^{o,k}\right) - g(\eta, P_h^o)| \leq H\sqrt{\frac{4S\log(3SAH^2K|N_\epsilon(\eta)|/\delta')}{N_h^k(s,a)}} + 2\epsilon\,,$$

where $|N_\epsilon(\eta)|$ is the size of the ϵ-net.

It now remains to bound the size of $|N_\epsilon(\eta)|$, which can be obtained easily if g is Lischitz. Notice that

$$\begin{aligned} |g(\tilde{\eta_1}, P_h^o) - g(\tilde{\eta_2}, P_h^o)| &\leq \sum_{s'} P_h^o(s' \mid s,a)|\tilde{\eta_1} - \tilde{\eta_2}| + |\tilde{\eta_1} - \tilde{\eta_2}| + \frac{|\tilde{\eta_1} - \tilde{\eta_2}|}{2}\rho \\ &= \frac{4+\rho}{2}|\tilde{\eta_1} - \tilde{\eta_2}|\,, \end{aligned}$$

where the first inequality is by the absolute inequality and $|(a)_+ - (b)_+| \leq |a-b|$.

Then g is a $\frac{4+\rho}{2}$-Lipschitz function over $\eta \in [0, H]$, thus combined with Lemma 11, we have $|N_\epsilon(\eta)| = O\left(\frac{4+\rho}{2\epsilon}\right)$. Hence, we have the following inequality happens with at least $1-\delta'$ probability:

$$\max_{\eta\in[0,H]} |g\left(\eta, \hat{P}_h^{o,k}\right) - g\left(\eta, P_h^o\right)| \leq H\sqrt{\frac{4S\log(3SAH^2K(4+\rho)/2\epsilon\delta')}{N_h^k(s,a)}} + 2\epsilon\,.$$

Take $\epsilon = \frac{1}{2\sqrt{K}}$, we have the following inequality happens with at least $1-\delta'$ probability:

$$\begin{aligned}\sigma_{\mathcal{P}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s,a) - \sigma_{\hat{\mathcal{P}}_h(s,a)}(\hat{V}_{h+1}^{\pi_k})(s,a) &\leq \max_{\eta\in[0,H]} |g\left(\eta, \hat{P}_h^{o,k}\right) - g\left(\eta, P_h^o\right)| \\ &\leq H\sqrt{\frac{4S\log(3SAH^2K^{3/2}(4+\rho)/\delta')}{N_h^k(s,a)}} + \frac{1}{\sqrt{K}}\,.\end{aligned}$$

□

E Proof of Theorem 2

E.1 Good events

We first define the following good events, in which case we estimate the reward function and the nominal transition functions fairly accurately.

$$\mathcal{G}_k^r = \left\{ \forall s,a,h : \left| r_h(s,a) - \hat{r}_h^k(s,a) \right| \le \sqrt{\frac{2\ln(2SAH^2K/\delta')}{N_h^k(s,a)}} \right\},$$
$$\mathcal{G}_k^p = \left\{ \forall s,a,h : \sigma_{\mathcal{P}_h(s)}(\hat{V}_{h+1}^{\pi_k})(s,a) - \sigma_{\hat{\mathcal{P}}_h(s)}(\hat{V}_{h+1}^{\pi_k})(s,a) \le C_h^k(s,a) \right\},$$

where

$$C_h^k(s,a) = AH\sqrt{\frac{4SA\log(3SA^2H^3K^{3/2}(4+\rho)/\delta')}{N_h^k(s,a)}} + \frac{1}{H\sqrt{K}}.$$

When the two good events happens at the same time, we say the algorithm in inside the good event $\mathcal{G} = \left(\bigcap_{k=1}^K \mathcal{G}_k^r\right) \bigcap \left(\bigcap_{k=1}^K \mathcal{G}_k^p\right)$. The following lemma shows that $\mathcal{G}$ happens with high probability.

Lemma 5 (Good event)**.** *Let $\delta = 2\delta'$, then the good event happens with high probability, i.e. $\mathbb{P}[\mathcal{G}] \ge 1-\delta$.*

Proof. By Hoeffding's inequality and an union bound on all s, a, all possible values of $N_k(s,a)$ and k, we have $\mathbb{P}\left[\bigcap_{k=1}^K \mathcal{G}_k^r\right] \ge 1-\delta'$. By Lemma 7, we have $\mathbb{P}\left[\bigcap_{k=1}^K \mathcal{G}_k^p\right] \ge 1-\delta'$ Then set $\delta = 2\delta'$ and we have the desired result. □

E.2 Design of the bonus function

In the case of s-rectangular uncertainty set, we use the following bonus function $b_h^k(s,a)$ to encourage exploration.

$$b_h^k(s,a) = AH\sqrt{\frac{4SA\log(3SA^2H^2K^{3/2}(4+\rho)/\delta)}{N_h^k(s,a)}} + \frac{1}{\sqrt{K}} + \sqrt{\frac{2\log(3SAH^2K/\delta')}{N_h^k(s,a)}}. \quad (5)$$

E.3 Regret analysis

Theorem 2 (Regret under s-rectangular uncertainty set)**.** *With learning rate $\beta = \sqrt{\frac{2\log A}{H^2K}}$ and bonus term b_h^k as (5), with probability at least $1-\delta$, the regret of Algorithm 1 is bounded by $O\left(SA^2H^2\sqrt{K\log(SA^2H^2K^{3/2}(1+\rho)/\delta)}\right)$.*

Proof. Similar to the case of (s,a)-rectangular set, we start with decomposing the regret as follows,

$$\begin{aligned}\text{Regret}(K) &= \sum_{k=1}^K V_1^*(s) - V_1^{\pi_k}(s) \\ &= \sum_{k=1}^K \left(V_1^*(s) - \hat{V}_1^{\pi_k}(s)\right) + \left(\hat{V}_1^{\pi_k}(s) - V_1^{\pi_k}(s)\right).\end{aligned}$$

By Lemma 2 and Lemma 6, with probability at least $1-\delta$, we have

$$\begin{aligned}\text{Regret}(K) &= O\left(H^2\sqrt{K\log A}\right) + O\left(SA^2H^2\sqrt{K\log(SA^2H^2K^{3/2}(1+\rho)/\delta)}\right) \\ &= O\left(SA^2H^2\sqrt{K\log(SA^2H^2K^{3/2}(1+\rho)/\delta)}\right).\end{aligned}$$

□

Lemma 6. *With Algorithm 1, we have*

$$\sum_{k=1}^{K}(\hat{V}_1^{\pi_k} - V_1^{\pi_k})(s) = O\left(SA^2H^2\sqrt{K\log(SA^2H^2K^{3/2}(1+\rho)/\delta)}\right).$$

Proof. Similar to the case with (s,a)-rectangular uncertainty set, for any k, we can decompose $(\hat{V}_1^{\pi_k} - \hat{V}_1^{\pi_k})(s)$ as,

$$\begin{aligned}
&(\hat{V}_1^{\pi_k} - \hat{V}_1^{\pi_k})(s) \\
&\leq \sum_{h=1}^{H}\mathbb{E}_{\pi_k,\{p_t\}_{t=1}^h}\left[(r_h^k(s,a) - \hat{r}_h^k(s,a)) + \left(\sigma_{\hat{\mathcal{P}}_h(s)}\left(\hat{V}_{h+1}^{\pi_k}\right)(s,a) - \sigma_{\mathcal{P}_h(s)}\left(\hat{V}_{h+1}^{\pi_k}\right)(s,a)\right) + b_h^k(s,a)\right].
\end{aligned}$$

Thus by the design of our bonus function and with probability at least $1-\delta$, we have

$$\begin{aligned}
&\sum_{k=1}^{K}(\hat{V}_1^{\pi_k} - V_1^{\pi_k})(s) \\
&\leq 2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi_k,\{p_t\}_{t=1}^h}\left[b_h^k(s,a)\right] \\
&= H\sqrt{K} + O\left(HA\sqrt{SA\log(SA^2H^2K^{3/2}(1+\rho)/\delta)}\right)\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi_k,\{p_t\}_{t=1}^h}\left[\sqrt{\frac{1}{N_h^k(s,a)}}\right].
\end{aligned}$$

By Lemma 12, we have the bound of visitation counts:

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{1}{N_h^k(s,a)}} \leq 2H\sqrt{SAK}.$$

Combining everything, conditioned on the good event we have

$$\sum_{k=1}^{K}(\hat{V}_1^{\pi_k} - V_1^{\pi_k})(s) = O\left(SA^2H^2\sqrt{K\log(SA^2H^2K^{3/2}(1+\rho)/\delta)}\right).$$

□

Lemma 7. *For any h,k,s,a, the following inequality holds with probability at least $1-\delta$,*

$$\sigma_{\hat{\mathcal{P}}_h(s)}(\hat{V}_{h+1}^{\pi_k})(s,a) - \sigma_{\mathcal{P}_h(s)}(\hat{V}_{h+1}^{\pi_k})(s,a) \leq AH\sqrt{\frac{4SA\log(3SA^2H^2K^{3/2}(4+\rho)/\delta)}{N_h^k(s,a)}} + \frac{1}{\sqrt{K}}.$$

Proof. By the definition of $\sigma_{\mathcal{P}_h(s)}(\hat{V}_{h+1}^{\pi_k})(s,a) = \inf_{P_h\in\mathcal{P}_h}\sum_{s'}P_h(s'\mid s,a)\hat{V}_{h+1}^{\pi_k}(s')$, we consider the following optimization problem:

$$\begin{aligned}
&\min_{P_h}\ \sum_{s'}P_h(s'\mid s,a)\hat{V}_{h+1}^{\pi_k}(s') \\
&\text{s.t.}\quad \begin{cases} \sum_{s',a'}|P_h(s'\mid s,a') - P_h^o(s'\mid s,a')| \leq A\rho, \\ \sum_{s'}P_h(s'\mid s,a') = 1, \forall a'\in\mathcal{A}, \\ P_h^o(\cdot\mid s,a') > 0, P_h(\cdot\mid s,a') \geq 0, \forall a'\in\mathcal{A}. \end{cases}
\end{aligned}$$

Let $\tilde{P}_h(s'\mid s,a) = \frac{P_h(s'|s,a)}{P_h^o(s'|s,a)}$, we can rewrite the above optimization problem as

$$\begin{aligned}
&\min_{\tilde{P}_h}\sum_{s'}\tilde{P}_h(s'\mid s,a)P_h^o(s'\mid s,a)\hat{V}_{h+1}^{\pi_k}(s') \\
&\text{s.t.}\quad \begin{cases} \sum_{s',a'}|(\tilde{P}_h(s'\mid s,a') - 1|P_h^o(s'\mid s,a') \leq A\rho, \\ \sum_{s'}\tilde{P}_h(s'\mid s,a')P_h^o(s'\mid s,a') = 1, \quad \forall a'\in\mathcal{A} \\ \tilde{P}_h(\cdot\mid s,a') \geq 0, \quad \forall a'\in\mathcal{A}. \end{cases}
\end{aligned}$$

Use the Lagrangian multiplier method and $f(x) = |x-1|$, we have the Lagrangian $L(\tilde{P}_h, \eta, \lambda)$ with multiplier $\eta = \{\eta_a\}_{a\in\mathcal{A}}, \eta_a \in \mathbb{R}$, $\lambda \geq 0$,

$$
\begin{aligned}
& L\left(\tilde{P}_h, \eta, \lambda\right)(s,a) \\
=& \sum_{s'} \tilde{P}_h(s' \mid s,a) P_h^o(s' \mid s,a) \hat{V}_{h+1}^{\pi_k}(s') + \lambda \left(\sum_{s',a'} \left| (\tilde{P}_h(s' \mid s,a') - 1 \right| P_h^o(s' \mid s,a') - A\rho \right) \\
& - \sum_{a'} \eta_{a'} \left(\sum_{s'} \tilde{P}_h(s' \mid s,a') P_h^o(s' \mid s,a') - 1 \right) \\
=& -\lambda A\rho + \sum_{a'} \eta_{a'} + \lambda \sum_{s',a'} P_h^o(s' \mid s,a') \left(f\left(\tilde{P}_h(s' \mid s,a')\right) - \tilde{P}_h(s' \mid s,a') \left(\frac{\eta_{a'} - \mathbb{I}\{a'=a\} V_{h+1}^{\pi_k}(s')}{\lambda} \right) \right).
\end{aligned}
$$

The convex conjugate of f is $f^*(y) = \max_x \langle x, y\rangle - f(x)$. Using f^*, we can thus optimize over $\tilde{P}_h$ and rewrite the Lagrangian over as

$$
\begin{aligned}
L(\eta,\lambda)(s,a) =& \min_{\tilde{P}_h} L\left(\tilde{P}_h, \eta, \lambda\right)(s,a) \\
=& -\lambda A\rho + \sum_{a'} \eta_{a'} - \lambda \sum_{s',a'} P_h^o(s' \mid s,a') f^* \left(\frac{\eta_{a'} - \mathbb{I}\{a'=a\} V_{h+1}^{\pi_k}(s')}{\lambda} \right).
\end{aligned}
$$

Conditioned on $x \geq 0$, $f(x) = |x-1|$, notice that the conjugate $f^*(y)$ has the following closed form,

$$
f^*(y) = \max_x \langle x, y\rangle - f(x) = \begin{cases} -1 & y \leq -1\,, \\ y & y \in [-1,1]\,, \\ +\infty & y > 1\,. \end{cases}
$$

Let $\tilde{\eta}_a = \eta_a + \lambda$, using the closed form of $f^*(y)$, the equality $\max\{a,b\} = (a-b)_+ + b$ and conditioned on $\frac{\eta_{a'} - \mathbb{I}\{a'=a\} V_{h+1}^{\pi_k}(s')}{\lambda} \leq 1$, we can rewrite the optimization problem as

$$
\begin{aligned}
L(\tilde{\eta},\lambda)(s,a) =& -\lambda A\rho + \sum_{a'} \eta_{a'} - \lambda \sum_{s',a'} P_h^o(s' \mid s,a') f^* \left(\frac{\eta_{a'} - \mathbb{I}\{a'=a\} V_{h+1}^{\pi_k}(s')}{\lambda} \right) \\
=& -\lambda A\rho - \lambda A + \sum_{a'} \tilde{\eta}_{a'} - \lambda \sum_{s',a'} P_h^o(s' \mid s,a') \max\left\{ \frac{\eta_{a'} - \mathbb{I}\{a'=a\} V_{h+1}^{\pi_k}(s')}{\lambda}, -1 \right\} \\
=& -\lambda A\rho + \sum_{a'} \tilde{\eta}_{a'} - \sum_{s',a'} P_h^o(s' \mid s,a') \left(\tilde{\eta}_{a'} - \mathbb{I}\{a'=a\} V_{h+1}^{\pi_k}(s') \right)_+ .
\end{aligned}
$$

where constraint of λ is

$$
\lambda \geq 0, \quad \tilde{\eta}_{a'} - \mathbb{I}\{a'=a\} V_{h+1}^{\pi_k}(s') \leq 2\lambda,\ \forall a', s'\,.
$$

Note that the above Lagrangian is inversely proportional to λ and it achieves the maximum when $\lambda = \max_{s',a'} \frac{(\tilde{\eta}_{a'} - \mathbb{I}\{a'=a\} V_{h+1}^{\pi_k}(s'))_+}{2}$. Directly optimize over λ, we can reduce the problem to

$$
L(\tilde{\eta})(s,a) = \sum_{a'} \tilde{\eta}_{a'} - \sum_{s',a'} P_h^o(s' \mid s,a') \left(\tilde{\eta}_{a'} - \mathbb{I}\{a'=a\} V_{h+1}^{\pi_k}(s') \right)_+ - \max_{s',a'} \frac{A\rho(\tilde{\eta}_{a'} - \mathbb{I}\{a'=a\} V_{h+1}^{\pi_k}(s'))_+}{2}\,.
$$

Define $g\left(\tilde{\eta}, P_h^o\right) = -L(\tilde{\eta})(s,a)$ as

$$
g(\tilde{\eta}, P_h^o) = -\sum_{a'} \tilde{\eta}_{a'} + \sum_{s',a'} P_h^o(s' \mid s,a') \left(\tilde{\eta}_{a'} - \mathbb{I}\{a'=a\} V_{h+1}^{\pi_k}(s') \right)_+ + \max_{s',a'} \frac{A\rho(\tilde{\eta}_{a'} - \mathbb{I}\{a'=a\} V_{h+1}^{\pi_k}(s'))_+}{2}\,.
$$

Assume g achieves its minimum when $\tilde{\eta} = \{\tilde{\eta}_1, \cdots, \tilde{\eta}_A\}$. Suppose $\tilde{\eta}$ has a component $\tilde{\eta}_a < 0$. Consider $\eta' = \{\tilde{\eta}_1, \cdots, 0, \cdots, \tilde{\eta}_a\}$, where we change the zero element $\tilde{\eta}_a$ to 0 and keep other components unchanged. Then we have

$$g(\tilde{\eta}, P_h^o) - g(\eta', P_h^o) = -\tilde{\eta}_A > 0\,,$$

which contradict with the hypothesis that g achieves its minimum in $\tilde{\eta}$.

On the other hand, suppose $\tilde{\eta}$ has a component $\tilde{\eta}_a > H$. Then consider $\eta' = \{\tilde{\eta}_1, \cdots, H, \cdots, \tilde{\eta}_a\}$, where we change corresponding $\tilde{\eta}_a$ to 0 and keep other components unchanged. Denote $f(\tilde{\eta}) = \max_{s',a'} \frac{A\rho(\tilde{\eta}_{a'} - \mathbb{I}\{a'=a\}V_{h+1}^{\pi_k}(s'))_+}{2}$, and we have

$$\begin{aligned} g\left(\tilde{\eta}, P_h^o\right) - g\left(\eta', P_h^o\right) =& -\tilde{\eta}_A + H + \sum_{s'} P_h^o(s' \mid s, a)(\tilde{\eta}_a - H) + f(\tilde{\eta}) - f(\eta') \\ \geq& -\tilde{\eta}_A + H + \sum_{s'} P_h^o(s' \mid s, a)(\tilde{\eta}_a - H) \\ =& 0\,. \end{aligned}$$

Therefore, g achieves its minimum with $\tilde{\eta}$, with $0 \leq \eta_a \leq H, \forall a \in \mathcal{A}$. We remark that a similar form and technique are also used for analyzing robust policy evaluation (Lemma C.1 [38]).

We can now rewrite

$$\begin{aligned} \sigma_{\hat{\mathcal{P}}_h(s)}\left(\hat{V}_{h+1}^{\pi_k}\right)(s,a) - \sigma_{\mathcal{P}_h(s)}\left(\hat{V}_{h+1}^{\pi_k}\right)(s,a) =& \min_{\eta_1 \in [0,H]^{|\mathcal{A}|}} g(\eta_1, \hat{P}_h^{o,k}) - \min_{\eta_2 \in [0,H]^{|\mathcal{A}|}} g(\eta_2, P_h^o) \\ \leq& \max_{\eta \in [0,H]^{|\mathcal{A}|}} \left| g\left(\eta, \hat{P}_h^{o,k}\right) - g\left(\eta, P_h^o\right) \right|\,. \end{aligned}$$

To upper bound $\sigma_{\hat{\mathcal{P}}_h(s)}\left(\hat{V}_{h+1}^{\pi_k}\right)(s,a) - \sigma_{\mathcal{P}_h(s)}\left(\hat{V}_{h+1}^{\pi_k}\right)(s,a)$, we first consider the bound of $\left|g\left(\eta, \hat{P}_h^{o,k}\right) - g\left(\eta, P_h^o\right)\right|$,

$$\begin{aligned} &\left|g\left(\eta, \hat{P}_h^{o,k}\right) - g\left(\eta, P_h^o\right)\right| \\ =& \left| \sum_{s',a'} \hat{P}_h^{o,k}(s' \mid s, a') \left(\eta_{a'} - \mathbb{I}\{a'=a\}V_{h+1}^{\pi_k}(s')\right)_+ - \sum_{s',a'} P_h^o(s' \mid s, a') \left(\eta_{a'} - \mathbb{I}\{a'=a\}V_{h+1}^{\pi_k}(s')\right)_+ \right| \\ =& \left| \sum_{a'} \sum_{s'} \left(\hat{P}_h^{o,k}(s' \mid s, a') - P_h^o(s' \mid s, a')\right) \left(\eta_{a'} - \mathbb{I}\{a'=a\}V_{h+1}^{\pi_k}(s')\right)_+ \right| \\ \leq& \sum_{a'} \left\| \hat{P}_h^{o,k}(\cdot \mid s, a') - P_h^o(\cdot \mid s, a') \right\|_1 \max_{s \in \mathcal{S}} \left| \eta_{a'} - \mathbb{I}\{a'=a\}V_{h+1}^{\pi_k}(s) \right| \\ \leq& H \sum_{a'} \left\| \hat{P}_h^{o,k}(\cdot \mid s, a') - P_h^o(\cdot \mid s, a') \right\|_1\,, \end{aligned}$$

where the first inequality is by Cauchy-Schwarz inequality, the second inequality follows from $\eta_a \in [0, H],\ \forall a \in \mathcal{A}$.

By Hoeffding's inequality and an union bound over all s, a', $N_h^k(s,a)$, the following inequality holds with probability at least $1-\delta$,

$$\left\| \hat{P}_h^{o,k}(\cdot \mid s, a') - P_h^o(\cdot \mid s, a') \right\|_1 \leq \sqrt{\frac{4S\log(SAH^2K/\delta)}{N_h^k(s,a)}}\,.$$

To upper bound $\max_{\eta \in [0,H]^{|\mathcal{A}|}} \left|g\left(\eta, \hat{P}_h^{o,k}\right) - g\left(\eta, P_h^o\right)\right|$, we first create an ϵ-net $N_\epsilon(\eta)$ with g over $\eta \in [0, H]$ such that

$$\max_{\eta \in [0,H]} \left|g\left(\eta, \hat{P}_h^{o,k}\right) - g\left(\eta, P_h^o\right)\right| \leq \max_{\eta \in N_\epsilon(\eta)} \left|g\left(\eta, \hat{P}_h^{o,k}\right) - g\left(\eta, P_h^o\right)\right| + 2\epsilon\,.$$

Taking an union bound over $N_\epsilon(\eta)$, we have

$$\max_{\eta\in[0,H]}\left|g\left(\eta,\hat{P}_h^{o,k}\right)-g\left(\eta,P_h^o\right)\right|\le HA\sqrt{\frac{4S\log(3SAH^2K|N_\epsilon(\eta)|/\delta)}{N_h^k(s,a)}}+2\epsilon\,,$$

where $|N_\epsilon(\eta)|$ is the size of the ϵ-net.

It now remains to find the size of the ϵ-net, which can be easily obtained if g is Lipschitz. Notice that

$$\begin{aligned}
&|g(\tilde{\eta}_1,P_h^o)-g(\tilde{\eta}_2,P_h^o)|\\
\le&\sum_{s',a'}P_h^o(s'\mid s,a)|\tilde{\eta}_{1,a'}-\tilde{\eta}_{2,a'}|+\sum_{a'}|\tilde{\eta}_{1,a'}-\tilde{\eta}_{2,a'}|+\frac{\max_{a'}|\tilde{\eta}_{1,a'}-\tilde{\eta}_{2,a'}|}{2}A\rho\\
\le&\frac{A(4+\rho)}{2}\|\tilde{\eta}_1-\tilde{\eta}_2\|_\infty\,,
\end{aligned}$$

where the first inequality is by the absolute inequality, the property of maximum function and $|(a)_+-(b)_+|\le|a-b|$, the second inequality follows from the definition of infinity norm.

Therefore g is a $\frac{A(4+\rho)}{2}$-Lipschitz function over $\eta\in[0,H]$. Thus combining with Lemma 11, we have $|N_\epsilon(\eta)|\le\left(\frac{A(4+\rho)}{2\epsilon}\right)^A$. Hence, we have the following inequality happens with at least $1-\delta'$ probability:

$$\begin{aligned}
\sigma_{\hat{\mathcal{P}}_h(s)}(\hat{V}_{h+1}^{\pi_k})(s,a)-\sigma_{\mathcal{P}_h(s)}(\hat{V}_{h+1}^{\pi_k})(s,a)\le&\max_{\eta_a\in[0,H]^{|\mathcal{A}|}}\left|g\left(\eta,\hat{P}_h^{o,k}\right)-g\left(\eta,P_h^o\right)\right|\\
\le&AH\sqrt{\frac{4SA\log(3SA^2H^2K(4+\rho)/2\epsilon\delta')}{N_h^k(s,a)}}+2\epsilon\,.
\end{aligned}$$

Take $\epsilon=\frac{1}{2\sqrt{K}}$, then

$$\sigma_{\hat{\mathcal{P}}_h(s)}(\hat{V}_{h+1}^{\pi_k})(s,a)-\sigma_{\mathcal{P}_h(s)}(\hat{V}_{h+1}^{\pi_k})(s,a)\le AH\sqrt{\frac{4SA\log(3SA^2H^2K^{3/2}(4+\rho)/\delta')}{N_h^k(s,a)}}+\frac{1}{\sqrt{K}}\,.$$

□

F Extension to uncertainty set with KL divergence

In this section, we extend our algorithm and analysis to uncertainty sets with KL divergence as a distance metric. We first formally define the uncertainty set considered, which is similar to the one in Definition 2.1.

Definition F.1 ((s,a)-rectangular uncertainty set Iyengar [12], Wiesemann et al. [36])**.** *For all time step h and with a given state-action pair (s,a), the (s,a)-rectangular uncertainty set $\mathcal{P}_h(s,a)$ is defined as*

$$\mathcal{P}_h(s,a) = \{D_{KL}\left(P_h(\cdot \mid s,a), P_h^o(\cdot \mid s,a)\right) \leq \rho\,, P_h(\cdot \mid s,a) \in \Delta(\mathcal{S})\}\,,$$

where P_h^o is the nominal transition kernel at h, $P_h^o(\cdot \mid s,a) > 0, \forall (s,a) \in \mathcal{S} \times \mathcal{A}$, ρ is the level of uncertainty and $D_{KL}\left(p(\cdot \mid s,a), q(\cdot \mid s,a)\right) = \sum_{s' \in \mathcal{S}} p(s' \mid s,a) \log\left(\frac{p(s'|s,a)}{q(s'|s,a)}\right)$.

With the above described uncertainty set, our algorithm solves $\sigma_{\hat{\mathcal{P}}_h}(\hat{V}_{h+1}^{\pi})(s,a)$ by solving the following sub-problem,

$$\min_{\lambda} \lambda\rho + \lambda \log\left(\sum_{s'} \hat{P}_h^o(s' \mid s,a) \exp\left(\frac{-\hat{V}_{h+1}^{\pi_k}(s')}{\lambda}\right)\right)\,.$$

Our algorithm also uses the following bonus function in the robust policy evaluation step,

$$b_h^k(s,a) = C_h^k(s,a) + \sqrt{\frac{2\log(3SAH^2K/\delta')}{N_h^k(s,a)}}\,.$$

With these modifications to algorithm 1, the following theorem states the formal regret guarantee.

Theorem 3 (Regret under KL divergence (s,a)-rectangular uncertainty set)**.** *Setting the learning rate $\beta = \sqrt{\frac{2\log A}{H^2K}}$, then with probability at least $1-\delta$, the regret incurred by Algorithm over K episodes is bounded by*

$$\textit{Regret}(K) = O\left(\frac{SH}{\rho c}\sqrt{AK\log(SAH^4K^{3/2}/\delta)}\right)\,,$$

where $0 < c \leq 1$ the minimal element of P_h^o, over all $h \in [H]$.

In the following, we present the detailed analysis of Theorem 3

F.1 Good events

We first define the following good events, in which case we estimate the reward function and the nominal transition functions fairly accurately.

$$\mathcal{G}_k^r = \left\{\forall s,a,h : \left|r_h(s,a) - \hat{r}_h^k(s,a)\right| \leq \sqrt{\frac{2\ln(2SAH^2K/\delta')}{N_h^k(s,a)}}\right\}\,,$$

$$\mathcal{G}_k^p = \left\{\forall s,a,h : \sigma_{\mathcal{P}_h(s)}(\hat{V}_{h+1}^{\pi_k})(s,a) - \sigma_{\hat{\mathcal{P}}_h(s)}(\hat{V}_{h+1}^{\pi_k})(s,a) \leq C_h^k(s,a)\right\}\,,$$

where

$$C_h^k(s,a) = \frac{2H}{\rho c}\sqrt{\frac{4S\log(8SAH^4K^2/\delta'\rho)}{N_h^k(s,a)}} + \frac{1}{\sqrt{K}}\,,$$

and c is the minimal element of P_h^o, over all $h \in [H]$. When the two good events happens at the same time, we say the algorithm in inside the good event $\mathcal{G} = \left(\bigcap_{k=1}^K \mathcal{G}_k^r\right) \bigcap \left(\bigcap_{k=1}^K \mathcal{G}_k^p\right)$. The following lemma shows that $\mathcal{G}$ happens with high probability.

Lemma 8 (Good event)**.** *Let $\delta = 2\delta'$, then the good event happens with high probability, i.e. $\mathbb{P}\left[\mathcal{G}\right] \geq 1-\delta$.*

Proof. By Hoeffding's inequality and an union bound on all s,a, all possible values of $N_k(s,a)$ and k, we have $\mathbb{P}\left[\bigcap_{k=1}^K \mathcal{G}_k^r\right] \geq 1-\delta'$. By Lemma 10, we have $\mathbb{P}\left[\bigcap_{k=1}^K \mathcal{G}_k^p\right] \geq 1-\delta'$ Then set $\delta = 2\delta'$ and we have the desired result. □

F.2 Regret analysis

Proof. Similar to the case of (s,a)-rectangular set, we start with decomposing the regret as follows,

$$\begin{aligned}\text{Regret}(K) &= \sum_{k=1}^{K} V_1^*(s) - V_1^{\pi_k}(s) \\ &= \sum_{k=1}^{K} \left(V_1^*(s) - \hat{V}_1^{\pi_k}(s)\right) + \left(\hat{V}_1^{\pi_k}(s) - V_1^{\pi_k}(s)\right) .\end{aligned}$$

By Lemma 2 and Lemma 9, with probability at least $1-\delta$, we have

$$\begin{aligned}\text{Regret}(K) &= O\left(H^2\sqrt{K\log A}\right) + O\left(\frac{SH}{\rho c}\sqrt{AK\log(SAH^4K^{3/2}/\delta)}\right) \\ &= O\left(\frac{SH}{\rho c}\sqrt{AK\log(SAH^4K^{3/2}/\delta)}\right) ,\end{aligned}$$

where c is the minimal element of P_h^o, over all $h\in[H]$. □

Lemma 9. *With Algorithm 1, we have*

$$\sum_{k=1}^{K}(\hat{V}_1^{\pi_k} - V_1^{\pi_k})(s) = O\left(\frac{1}{\rho c}SH\sqrt{AK\log(SAH^4K^{3/2}/\delta)}\right) .$$

Proof. Similar to the case with (s,a)-rectangular uncertainty set, for any k, we can decompose $(\hat{V}_1^{\pi_k} - \hat{V}_1^{\pi_k})(s)$ as,

$$(\hat{V}_1^{\pi_k} - \hat{V}_1^{\pi_k})(s) \le \sum_{h=1}^{H}\mathbb{E}_{\pi_k,\{p_t\}_{t=1}^h}\left[(r_h^k(s,a) - \hat{r}_h^k(s,a)) + \left(\sigma_{\hat{\mathcal{P}}_h(s)}\left(\hat{V}_{h+1}^{\pi_k}\right)(s,a) - \sigma_{\mathcal{P}_h(s)}\left(\hat{V}_{h+1}^{\pi_k}\right)(s,a)\right) + b_h^k(s,a)\right] .$$

Thus by the design of our bonus function and with probability at least $1-\delta$, we have

$$\begin{aligned}&\sum_{k=1}^{K}(\hat{V}_1^{\pi_k} - V_1^{\pi_k})(s) \\ &\le 2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi_k,\{p_t\}_{t=1}^h}\left[b_h^k(s,a)\right] \\ &= H\sqrt{K} + O\left(\frac{1}{\rho c}\sqrt{S\log(SAH^4K^{3/2}/\delta)}\right)\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi_k,\{p_t\}_{t=1}^h}\left[\sqrt{\frac{1}{N_h^k(s,a)}}\right] ,\end{aligned}$$

where c is a problem dependent constant.

By Lemma 12, we have the bound of visitation counts:

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{1}{N_h^k(s,a)}} \le 2H\sqrt{SAK} .$$

Combining everything, conditioned on the good event we have

$$\sum_{k=1}^{K}(\hat{V}_1^{\pi_k} - V_1^{\pi_k})(s) = O\left(\frac{SH}{\rho c}\sqrt{AK\log(SAH^4K^{3/2}/\delta)}\right) .$$

□

Lemma 10. *For any h,k,s,a, the following inequality holds with probability at least $1-\delta$,*

$$\sigma_{\hat{\mathcal{P}}_h(s)}(\hat{V}_{h+1}^{\pi_k})(s,a) - \sigma_{\mathcal{P}_h(s)}(\hat{V}_{h+1}^{\pi_k})(s,a) \le \frac{2H}{\rho c}\sqrt{\frac{4S\log(8SAH^4K^2/\delta'\rho)}{N_h^k(s,a)}} + \frac{1}{\sqrt{K}} .$$

where c is the minimal element of P_h^o.

Proof. By the definition of $\sigma_{\mathcal{P}_h(s)}\left(\hat{V}_{h+1}^{\pi_k}\right)(s,a) = \inf_{P_h \in \mathcal{P}_h} \sum_{s'} P_h(s' \mid s,a)\hat{V}_{h+1}^{\pi_k}(s')$, we consider the following optimization problem:

$$\min_{P_h} \sum_{s'} P_h(s' \mid s,a)\hat{V}_{h+1}^{\pi_k}(s')$$
$$\text{s.t.} \quad \begin{cases} \sum_{s'} P_h(s' \mid s,a) \log\left(\frac{P_h(s'|s,a)}{P_h^o(s'|s,a)}\right) \le \rho\,, \\ \sum_{s'} P_h(s' \mid s,a) = 1\,, \\ P_h^o(\cdot \mid s,a) > 0, P_h(\cdot \mid s,a) \ge 0\,. \end{cases}$$

Let $\tilde{P}_h(s' \mid s,a) = \frac{P_h(s'|s,a)}{P_h^o(s'|s,a)}$, we can rewrite the above optimization problem as

$$\min_{\tilde{P}_h} \sum_{s'} \tilde{P}_h(s' \mid s,a)P_h^o(s' \mid s,a)\hat{V}_{h+1}^{\pi_k}(s')$$
$$\text{s.t.} \quad \begin{cases} \sum_{s'} \tilde{P}_h(s' \mid s,a')P_h^o(s' \mid s,a') \log\left(\tilde{P}_h(s' \mid s,a)\right) \le \rho\,, \\ \sum_{s'} \tilde{P}_h(s' \mid s,a')P_h^o(s' \mid s,a) = 1\,, \\ \tilde{P}_h(\cdot \mid s,a) \ge 0\,. \end{cases}$$

Use the Lagrangian multiplier method and $f(x) = x\log x$, we have the Lagrangian $L(\tilde{P}_h, \eta, \lambda)$ with multiplier $\eta \in \mathbb{R}$, $\lambda \ge 0$,

$$\begin{aligned} & L(\tilde{P}_h, \eta, \lambda)(s,a) \\ =& \sum_{s'} \tilde{P}_h(s' \mid s,a)P_h^o(s' \mid s,a)\hat{V}_{h+1}^{\pi_k}(s') + \lambda\left(\sum_{s'} \tilde{P}_h(s' \mid s,a')P_h^o(s' \mid s,a')\log(\tilde{P}_h(s' \mid s,a)) - \rho\right) \\ & - \eta\left(\sum_{s'} \tilde{P}_h(s' \mid s,a)P_h^o(s' \mid s,a) - 1\right) \\ =& -\lambda\rho + \eta + \lambda\sum_{s'} P_h^o(s' \mid s,a)\left(f\left(\tilde{P}_h(s' \mid s,a')\right) - \tilde{P}_h(s' \mid s,a')\left(\frac{\eta - V_{h+1}^{\pi_k}(s')}{\lambda}\right)\right)\,. \end{aligned}$$

The convex conjugate of f is $f^*(y) = \max_x \langle x,y\rangle - f(x)$. Using f^*, we can thus optimize over $\tilde{P}_h$ and rewrite the Lagrangian over as

$$L(\eta,\lambda)(s,a) = \min_{\tilde{P}_h} L(\tilde{P}_h,\eta,\lambda)(s,a) = -\lambda\rho + \eta - \lambda\sum_{s'} P_h^o(s' \mid s,a)f^*\left(\frac{\eta - V_{h+1}^{\pi_k}(s')}{\lambda}\right)\,.$$

Conditioned on $x \ge 0$, $f(x) = x\log x$, notice that the conjugate $f^*(y)$ has the following closed form,

$$f^*(y) = \max_x \langle x,y\rangle - f(x) = \exp(y-1)\,.$$

Using the closed form of $f^*(y)$, we can rewrite the optimization problem as

$$\begin{aligned} L(\eta,\lambda)(s,a) &= -\lambda\rho + \eta - \lambda\sum_{s'} P_h^o(s' \mid s,a)f^*\left(\frac{\eta - V_{h+1}^{\pi_k}(s')}{\lambda}\right) \\ &= -\lambda\rho + \eta - \lambda\sum_{s'} P_h^o(s' \mid s,a)\exp\left(\frac{\eta - V_{h+1}^{\pi_k}(s') - \lambda}{\lambda}\right)\,. \end{aligned}$$

Taking the derivative of η,

$$\frac{\partial L}{\partial \eta} = 1 - \sum_{s'} P_h^o(s' \mid s,a)\exp\left(\frac{\eta - V_{h+1}^{\pi_k}(s') - \lambda}{\lambda}\right) = 0\,,$$

$$\eta = \lambda - \lambda \log \left(\sum_{s'} P_h^o(s' \mid s,a) \exp\left(\frac{-V_{h+1}^{\pi_k}(s')}{\lambda} \right) \right).$$

Directly optimize over η, we can reduce the problem to

$$\begin{aligned} L(\lambda)(s,a) &= \lambda(1-\rho) - \lambda \log \left(\sum_{s'} P_h^o(s' \mid s,a) \exp\left(\frac{-V_{h+1}^{\pi_k}(s')}{\lambda} \right) \right) - \lambda \,, \\ &= -\lambda\rho - \lambda \log \left(\sum_{s'} P_h^o(s' \mid s,a) \exp\left(\frac{-V_{h+1}^{\pi_k}(s')}{\lambda} \right) \right). \end{aligned}$$

Define $g(\lambda, P_h^o) = -L(\lambda)(s,a)$ as

$$g(\lambda, P_h^o) = \lambda\rho + \lambda \log \left(\sum_{s'} P_h^o(s' \mid s,a) \exp\left(\frac{-V_{h+1}^{\pi_k}(s')}{\lambda} \right) \right).$$

Note that the Lagrangian multiplier $\lambda \geq 0$. Then we prove g is bounded within $[-H, H]$ over $[0, H/\rho]$.

$$\begin{aligned} g(\lambda, P_h^o) &= \lambda\rho + \lambda \log \left(\sum_{s'} P_h^o(s' \mid s,a) \exp\left(\frac{-V_{h+1}^{\pi_k}(s')}{\lambda} \right) \right), \\ &\leq \lambda\rho + \lambda \log \left(\sum_{s'} P_h^o(s' \mid s,a) \exp\left(\frac{-0}{\lambda} \right) \right), \\ &= \lambda\rho \leq H \,, \end{aligned}$$

where the first inequality follows from $V_{h+1}^{\pi_k}(s') \geq 0$ and the second inequality is by $\lambda \leq H/\rho$.

$$\begin{aligned} g(\lambda, P_h^o) &= \lambda\rho + \lambda \log \left(\sum_{s'} P_h^o(s' \mid s,a) \exp\left(\frac{-V_{h+1}^{\pi_k}(s')}{\lambda} \right) \right), \\ &\geq \lambda\rho + \lambda \log \left(\sum_{s'} P_h^o(s' \mid s,a) \exp\left(\frac{-H}{\lambda} \right) \right), \\ &= \lambda\rho - H \geq -H \,, \end{aligned}$$

where the first inequality follows from $V_{h+1}^{\pi_k}(s') \leq H$ and the second inequality is by $\lambda \geq 0$.

Moreover, from the induction above we know that for any P, $g(0,P) \leq 0$ and for $\lambda > H/\rho$,

$$g(\lambda, P) \geq \lambda\rho + \lambda \log(\exp(-H/\lambda)) > 0 \,.$$

Therefore, g achieves its minimum over $\lambda \in [0, H/\rho]$. We remark that the same form is also used for sample complexity results ([2, 38]).

We can now rewrite

$$\begin{aligned} \sigma_{\hat{\mathcal{P}}_h(s)}\left(\hat{V}_{h+1}^{\pi_k}\right)(s,a) - \sigma_{\mathcal{P}_h(s)}\left(\hat{V}_{h+1}^{\pi_k}\right)(s,a) &= \min_{0\leq\lambda_1\leq H/\rho} g\left(\lambda_1, \hat{P}_h^{o,k}\right) - \min_{0\leq\lambda_2\leq H/\rho} g\left(\lambda_2, P_h^o\right) \\ &\leq \max_{0\leq\lambda\leq H/\rho} \left| g\left(\lambda, \hat{P}_h^{o,k}\right) - g\left(\lambda, P_h^o\right) \right| . \end{aligned}$$

By [21] (Appendix C), when $\lambda = 0$, $g\left(\lambda, \hat{P}_h^{o,k}\right) = g\left(\lambda, P_h^o\right) = \min_{s\in\mathcal{S}} V_{h+1}^{\pi_k}(s)$. Therefore, it suffice to bound over $\max_{c\leq\lambda\leq H/\rho} \left| g\left(\lambda, \hat{P}_h^{o,k}\right) - g\left(\lambda, P_h^o\right) \right|$, where $c > 0$. We now have

$$\left| g\left(\lambda, \hat{P}_h^{o,k}\right) - g\left(\lambda, P_h^o\right) \right|$$

$$
\begin{aligned}
&= \left| \lambda \log \left(\sum_{s'} \hat{P}_h^{o,k}(s' \mid s,a) \exp\left(\frac{-V_{h+1}^{\pi_k}(s')}{\lambda} \right) \right) - \lambda \log \left(\sum_{s'} P_h^o(s' \mid s,a) \exp\left(\frac{-V_{h+1}^{\pi_k}(s')}{\lambda} \right) \right) \right| \\
&= \left| \lambda \log \left(1 + \frac{\sum_{s'} (\hat{P}_h^{o,k}(s' \mid s,a) - P_h^o(s' \mid s,a)) \exp\left(\frac{-V_{h+1}^{\pi_k}(s')}{\lambda} \right)}{\sum_{s'} P_h^o(s' \mid s,a) \exp\left(\frac{-V_{h+1}^{\pi_k}(s')}{\lambda} \right)} \right) \right| \\
&\leq 2\lambda \left| \frac{\sum_{s'} (\hat{P}_h^{o,k}(s' \mid s,a) - P_h^o(s' \mid s,a)) \exp\left(\frac{-V_{h+1}^{\pi_k}(s')}{\lambda} \right)}{\sum_{s'} P_h^o(s' \mid s,a) \exp\left(\frac{-V_{h+1}^{\pi_k}(s')}{\lambda} \right)} \right| \\
&\leq 2\lambda \max_{s'} \left| \frac{\hat{P}_h^{o,k}(s' \mid s,a) - P_h^o(s' \mid s,a)}{P_h^o(s' \mid s,a)} \right|
\end{aligned}
$$

where the first inequality follows from $|\log(1+x)| \leq 2|x|$ and the second inequality follows from the Holder's inequality.

By Hoeffding's inequality and an union bound over all s, a', $N_h^k(s,a)$, the following inequality holds with probability at least $1-\delta$,

$$
\max_{s'} \left| \hat{P}_h^{o,k}(s' \mid s,a) - P_h^o(s' \mid s,a) \right| \leq \left\| \hat{P}_h^{o,k}(\cdot \mid s,a) - P_h^o(\cdot \mid s,a) \right\|_1 \leq \sqrt{\frac{4S \log(SAH^2K/\delta)}{N_h^k(s,a)}} \,.
$$

Then we create an ϵ-net $N_\epsilon(\lambda)$ with g over $\lambda \in [0, H/\rho]$ such that

$$
\max_{\lambda \in [0,H/\rho]} |g(\lambda, \hat{P}_h^{o,k}) - g(\lambda, P_h^o)| \leq \max_{\lambda \in N_\epsilon(\eta)} |g(\lambda, \hat{P}_h^{o,k}) - g(\lambda, P_h^o)| + 2\epsilon \,.
$$

Then we know that $|N_\epsilon(\lambda)|$ is bounded by the area of the rectangle $[0, H/\rho] \times [-H, H]$ over ϵ^2,

$$
|N_\epsilon(\lambda)| \leq \frac{2H^2}{\rho \epsilon^2} \,.
$$

Taking an union bound over $N_\epsilon(\lambda)$ and denote $c = \min_{s'} P_h^o(\cdot \mid s,a)$, we have the following inequality happens with at least $1-\delta'$ probability:

$$
\begin{aligned}
\sigma_{\hat{\mathcal{P}}_h(s)}(\hat{V}_{h+1}^{\pi_k})(s,a) - \sigma_{\mathcal{P}_h(s)}(\hat{V}_{h+1}^{\pi_k})(s,a) &\leq \max_{\lambda \in [0,H/\rho]} |g(\lambda, \hat{P}_h^{o,k}) - g(\lambda, P_h^o)| \\
&\leq \max_{\lambda \in N_\epsilon(\lambda)} |g(\lambda, \hat{P}_h^{o,k}) - g(\lambda, P_h^o)| + 2\epsilon \\
&\leq 2\frac{H}{\rho} \max_{s'} \left| \frac{\hat{P}_h^{o,k}(s' \mid s,a) - P_h^o(s' \mid s,a)}{P_h^o(s' \mid s,a)} \right| + 2\epsilon \\
&\leq 2\frac{H}{\rho c} \sqrt{\frac{4S \log(2SAH^4K/\delta'\rho\epsilon^2)}{N_h^k(s,a)}} + 2\epsilon \,,
\end{aligned}
$$

Take $\epsilon = \frac{1}{2\sqrt{K}}$, then

$$
\sigma_{\hat{\mathcal{P}}_h(s)}(\hat{V}_{h+1}^{\pi_k})(s,a) - \sigma_{\mathcal{P}_h(s)}(\hat{V}_{h+1}^{\pi_k})(s,a) \leq 2\frac{H}{\rho c} \sqrt{\frac{4S \log(8SAH^4K^2/\delta'\rho)}{N_h^k(s,a)}} + \frac{1}{\sqrt{K}} \,.
$$

□

G Proof of Proposition 1

Claim A.1 (Suboptimality of non-robust optimal policy)**.** *There exists a robust MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, r, H \rangle$ with uncertainty set $\mathcal{P}$ of uncertainty radius ρ, such that the non-robust optimal policy is $\Omega(1)$-suboptimal to the uniformly random policy.*

Proof. We consider a robust MDP with three states s_0, s_1, s_2 and two actions a_0, a_1. Without loss of generality, we let s_0 be the initial state. On the initial state s_0, both actions will lead to a reward of 0. On state s_1, a reward of $1/(H-1)$ is given for both actions. On state s_2, a reward of $-1/(H-1)$ is given for both actions. The nominal transition dynamic of the MDP is the following. Taking action a_0 on s_0 will be transited to s_1 with a probability of ϵ and be transited to s_2 with a probability of ϵ, while $\epsilon > 0.5$. Taking the other action a_1 will have equal probability of transiting to s_1 and s_2. The states s_1 and s_2 are absorbing, in the sense that taking any action on these two states will be transited by to the same state. The transition of the MDP is also illustrated in Figure 3, where a dashed line denotes a probabilistic transition and a solid line denotes deterministic transition. With the nominal

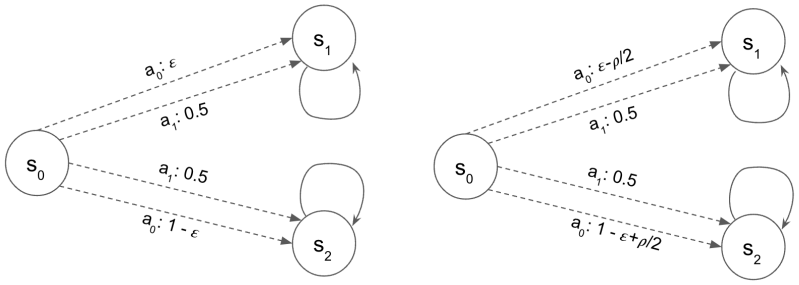


Figure 3: The left figure describes the nominal transition dynamic of the MDP. The right figure describes the robust transition dynamic of the MDP.

transition, it is clear that an optimal policy would be always taking a_0. Denote this policy as $\pi_{o,*}$, the value for this policy under nominal transition over K episodes is

$$V^{\pi_{o,*}}(s_0) = K(H-1)\left(\epsilon \cdot \frac{1}{H-1} - (1-\epsilon)\cdot \frac{1}{H-1}\right) = 2\epsilon - 1 > 0\,,$$

where the last inequality is due to $\epsilon > 0$.

However, consider the uncertainty radius ρ and the robust transition denoted by the right figure of Figure 3. That is, taking a_0 on s_0 will leads to a transition to s_1 with probability $\epsilon - \rho/2$ and to s_2 with probability $1 - \epsilon + \rho/2$. Note that as $\epsilon > 0.5$, $\rho \leq 1$, $\epsilon - \rho/2 > 0$. Moreover, this transition is indeed the worst case transition for any non-uniform policy. Let $\tilde{V}$ denotes the robust value under the above described transition. With a uniform policy π, the value of it under this transition is

$$\tilde{V}^{\pi}(s_0) = K(H-1)\left(0.5\left(\epsilon - \frac{\rho}{2}\right)\cdot \frac{1}{H-1} - 0.5\left(1-\epsilon+\frac{\rho}{2}\right))\cdot \frac{1}{H-1}\right) = \epsilon - \rho/2 - 0.5\,.$$

The value of $\pi_{o,*}$ is, however,

$$\tilde{V}^{\pi_{o,*}}(s_0) = K(H-1)\left(\left(\epsilon - \frac{\rho}{2}\right)\cdot \frac{1}{H-1} - \left(1-\epsilon+\frac{\rho}{2}\right))\cdot \frac{1}{H-1}\right) = 2\epsilon - \rho - 1\,.$$

For any $2\epsilon - 1 \leq \rho \leq 1$, we have $\tilde{V}^{\pi_{o,*}}(s_0) \leq \tilde{V}^{\pi}(s_0)$. Since $\epsilon > 0.5$ is arbitrary, the optimal policy under the nominal transition is non-robust even under the slightest perturbation. □

H Auxiliary lemmas

Lemma 11 ([3]). *An ϵ-cover of a subset T of a pseudometric space (S, d) is a set $\hat{T} \subset T$ such that for each $t \in T$ there is a $\hat{t} \in \hat{T}$ such that $d(t, \hat{t}) \leq \epsilon$. The ϵ-covering number of T is*

$$N(\epsilon, T, d) = \min \left\{ |\hat{T}| : \hat{T} \text{ is an } \epsilon\text{-cover of } T \right\} .$$

Let F_d be the set of L-Lipschitz functions (wrt $\|\cdot\|_\infty$) mapping from $[0,1]^d$ to $[0,1]$. Then

$$\log N\left(\epsilon, F_d, \|\cdot\|_\infty\right) = \Theta\left(\left(\frac{L}{\epsilon}\right)^d\right) .$$

Lemma 12 (Lemma 7.5 [1]). *For arbitrary K sequence of trajectories $\{s_h^k, a_h^k\}_{h=1}^H$, $k = 1, \ldots, K$, we have*

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \frac{1}{\sqrt{N_h^k(s_h^k, a_h^k)}} \leq 2H\sqrt{SAK} .$$

Proof. We have

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H} \frac{1}{\sqrt{N_h^k\left(s_h^k, a_h^k\right)}} &= \sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{i=1}^{N_h^K(s,a)} \frac{1}{\sqrt{i}} \\
&\leq 2\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sqrt{N_h^K(s,a)} \\
&\leq \sum_{h=1}^{H} \sqrt{SA\sum_{s,a} N_h^K(s,a)} \\
&= H\sqrt{SAK} ,
\end{aligned}$$

where the first inequality is by $\sum_{i=1}^{N} \frac{1}{\sqrt{i}} \leq 2\sqrt{N}$ and the second inequality follows by Cauchy-Schwarz inequality. □

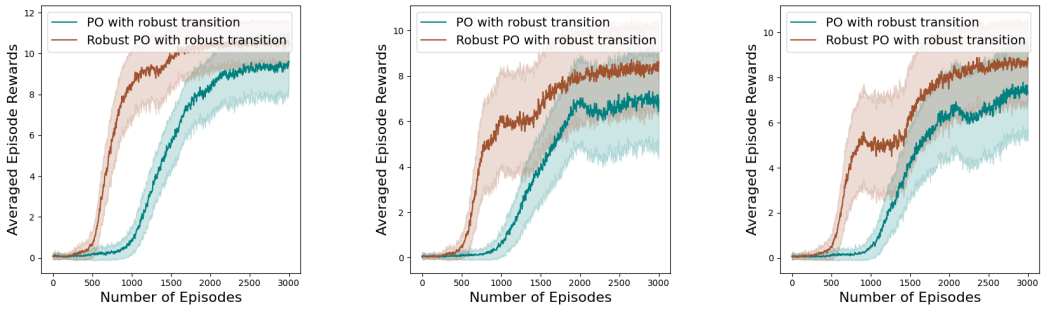
Lemma 13 (Fundamental inequality of Online Mirror Descent for RL (Lemma 17 [30])). *Let $\beta > 0$. Let $\pi_h^1(\cdot \mid s)$ be the uniform distribution. Then, by updating with OMD and with KL divergence regularization, for any $k \in [K], h \in [H]$ and $s \in \mathcal{S}$, the following holds for any stationary policy π,*

$$\sum_{k=1}^{K} \left\langle Q_h^k(\cdot \mid s), \pi_h^k(\cdot \mid s) - \pi_h(\cdot \mid s) \right\rangle \leq \frac{\log A}{\beta} + \frac{\beta}{2}\sum_{k=1}^{K}\sum_{a} \pi_h^k(a \mid s)\left(Q_h^k(s,a)\right)^2 . \tag{6}$$

I More experimental details

Other configurations and set up The episode length is set to 20 and all algorithms are trained with 3000 episodes. The evaluation results are averaged over 20 runs and is presented with 1 standard deviation. All experiments are conducted with 64 core ADM 3990X.

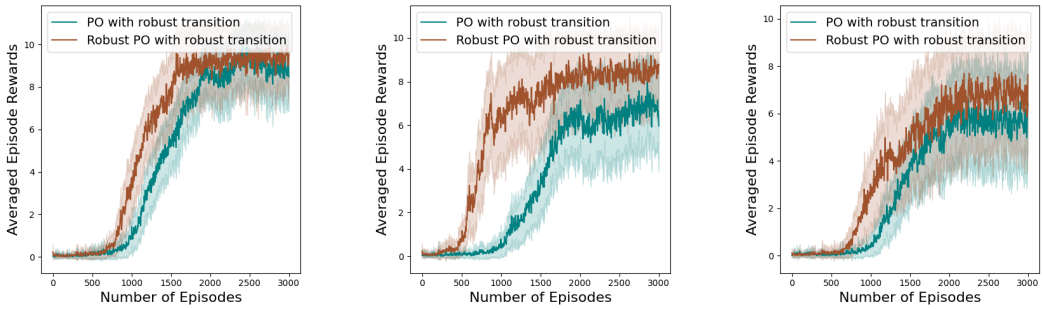
Results with ℓ_1 distance constrained s-rectangular uncertainty sets With the uncertainty set described with ℓ_1 distance with s-rectangular set, we present the following experimental results.



(a) $\rho = 0.1$ **(b)** $\rho = 0.2$ **(c)** $\rho = 0.3$

Figure 4: Cumulative rewards obtained by robust and non-robust policy optimization on robust transition with different level of uncertainty $\rho = 0.1, 0.2, 0.3$ under ℓ_1 distance, s-rectangular set.

Results with KL divergence constrained (s, a)-rectangular uncertainty sets With the uncertainty set described with KL divergence, we present the following experimental results. All other configurations and set up remains the same with those for uncertainty set with ℓ_1 distance.



(a) $\rho = 0.1$ **(b)** $\rho = 0.2$ **(c)** $\rho = 0.3$

Figure 5: Cumulative rewards obtained by robust and non-robust policy optimization on robust transition with different level of uncertainty $\rho = 0.1, 0.2, 0.3$ under KL divergence.