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 $\beta = \sqrt{\frac{2\log A}{H^2 K}}$ 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 $O\left(H^2 S\sqrt{AK\log\left(SAH^2K^{3/2}(1+\rho)/\delta\right)}\right)$.*

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.