Towards Faster Global Convergence of Robust Policy Gradient Methods

Anonymous authors
Paper under double-blind review

Abstract

We establish the global convergence of the policy gradient method for robust Markov Decision Processes (MDPs) under the assumption that the robust return is smooth with respect to a policy. Despite restrictive, such smoothness assumption is satisfied in many interesting settings such as reward-robust MDPs. We also obtain iteration complexity comparable to non-robust MDPs that is significantly faster than existing rates for robust MDPs.

1 Introduction

Robust MDPs model decision making problems in which the environment is uncertain or partially known (Nilim & Ghaoui, 2005; Iyengar, 2005a). The robust optimal policies are not only resilient to the changes in the environmental parameters, but also better generalizable (Mannor et al., 2007; Xu & Mannor, 2010). Most of the work on the topic are value-based methods (Lutter et al., 2021; Pinto et al., 2017; Ho et al., 2020; Badrinath & Kalathil, 2021; Behzadian et al., 2021; Goyal & Grand-Clément, 2018; Mannor et al., 2016; Wang et al., 2022c;a; Wang & Zou, 2021; Kumar et al., 2022), and only few are policy-based methods (Wang & Zou, 2022b; Kumar et al., 2023; Wang et al., 2022b). Moreover, up to our knowledge, there are only two works that establish global convergence of policy gradient methods: with iteration complexity of $O(SA\epsilon^{-3})$ for R-contamination uncertainty sets (Wang & Zou, 2022a) and $O(S^4A^2\epsilon^{-4})$ for general kernel uncertainty sets Wang et al. (2022b), where S and A are the number of states and the number of actions respectively, and ϵ is the target difference between optimal robust return and our achieved robust return.

The state-of-the-art iteration complexity for global convergence of policy gradient for non-robust MDPs is $O(SA\epsilon^{-1})$, however the proofs are quite technical and complicated (Agarwal et al., 2020; Mei et al., 2020; Xiao, 2022). Robustness requirement adds another layer of complexity to the problem, leading to a significant degradation in the iteration complexity (Wang & Zou, 2022a; Wang et al., 2022b).

We simplify and extend the machinery for non-robust MDPs (Xiao, 2022) to the robust MDPs, and achieve the same iteration complexity as for non-robust MDPs, i.e., $O(SA\epsilon^{-1})$. This complexity is achieved under the assumption that the robust return is L-smooth which holds in many useful scenarios (Gadot et al., 2023) (see Appendix B.2 for more examples).

Table 1: Iteration Complexity for Global Convergence of Policy Gradient Methods

Robust MDPs	Complexity	Remark
Non-Robust	$O(SA\epsilon^{-1})$	Xiao (2022)
(s,a) rectangular R-Contamination robust MDPs	$O(SA\epsilon^{-3})$	Wang & Zou (2022a)
General L-smooth robust MDPs	$O(SL\epsilon^{-1})$	Ours
General kernel robust MDPs	$O(S^4A^2\epsilon^{-4})$	Wang et al. (2022b)

2 Main

A robust Markov decision process (MDP) is a tuple $(\mathcal{S},\mathcal{A},\gamma,\mu,\mathcal{U})$ such that $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces respectively, $\gamma\in[0,1)$ is a discount factor and $\mu\in\Delta_{\mathcal{S}}$ the initial state distribution. Denoting $\mathcal{X} := \mathcal{S} \times \mathcal{A}$, the couple $(P, R) \in \mathcal{U}$ corresponds to the MDP model with $P : \mathcal{X} \to \Delta_{\mathcal{S}}$ being a transition kernel and $R : \mathcal{X} \to \mathbb{R}$ a reward function. A policy $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$ maps each state to a probability distribution over $\mathcal{A}$, and we denote by Π the set of possible policies. For any policy $\pi \in \Pi$, $R^\pi \in \mathbb{R}^{\mathcal{S}}$ is the expected immediate reward defined by $R^\pi(s) := \langle \pi_s, R(s, \cdot) \rangle_{\mathcal{A}}, \quad \forall s \in \mathcal{S}$, where π_s is a shorthand for $\pi(\cdot|s)$. We similarly define the stochastic matrix induced by π as $P^\pi(s'|s) := \langle \pi_s, P(s'|s, \cdot) \rangle_{\mathcal{A}}, \quad \forall s, s' \in \mathcal{S}$. Our goal is to maximize the robust discounted return $\min_{\pi \in \Pi} \rho^\pi_{\mathcal{U}}$ over the set of policies Π, with

$$\rho^\pi_{\mathcal{U}} := \min_{(P,R) \in \mathcal{U}} \rho^\pi_{(P,R)}, \quad \text{where return } \rho^\pi_{(P,R)} := \mu^T (I - \gamma P^\pi)^{-1} R^\pi. \tag{1}$$

However, the above optimization is proven to be Strongly NP-Hard for general convex uncertainty set (Wolfram Wiesemann, 2012). Hereafter, we assume the uncertainty set $\mathcal{U}$ to be convex, that allows robust return $\rho^\pi_{\mathcal{U}}$ to be differentiable and L-smooth w.r.t. policy π with $L > 0$, that is, $\left| \rho^{\pi'}_{\mathcal{U}} - \rho^{\pi'}_{\mathcal{U}} - \langle \nabla_\pi \rho^\pi_{\mathcal{U}}, \pi' - \pi \rangle \right| \leq \frac{L}{2} \|\pi' - \pi\|^2, \ \forall \pi', \pi \in \Pi$. In particular, the L_p-ball reward uncertainty set guarantees smoothness of the robust return (Gadot et al., 2023), more cases are discussed in Appendix B.2. To optimize the robust return, we rely on the projected gradient ascent: $\pi_{k+1} := \mathbf{proj}_\Pi(\pi_k + \frac{1}{L} \nabla_\pi \rho^{\pi_k}_{\mathcal{U}})$, where $\frac{1}{L}$ is the learning rate and $\mathbf{proj}_\Pi$ denotes the orthogonal projection onto set Π. For this policy gradient procedure, we get the following convergence result:

Theorem 1 (**Global optimality**). *For all iterations $k \geq 1$, it holds that:*

$$\rho^*_{\mathcal{U}} - \rho^{\pi_k}_{\mathcal{U}} \leq \frac{8 L C^2_{\text{PL}} \mathbf{diam}(\Pi)^2 (\rho^*_{\mathcal{U}} - \rho^{\pi_0}_{\mathcal{U}})}{k},$$

where $C_{\text{PL}} \leq \frac{1}{\min_s \mu(s)}$ is a problem dependent constant, that we define properly in Appendix C.3.

Proof. For the full proof see Appendix C. Below, we briefly describe the proof sketch. The main challenge of the proof is to extend the Gradient Domination property to the robust problem (sometimes called PL condition). This condition was proven to hold for non-robust MDPs (Xiao, 2022). Despite it might not hold for general robust problems, we prove it for smooth robust problems:

Lemma 1 (**Gradient Domination lemma**). *For any policy $\pi \in \Pi$, its sub-optimality is bounded by its policy gradient as $\rho^*_{\mathcal{U}} - \rho^\pi_{\mathcal{U}} \leq C_{\text{PL}} \max_{\pi' \in \Pi} \langle \pi' - \pi, \nabla \rho^\pi_{\mathcal{U}} \rangle$.*

For the proof see Appendix C.3. The rest of the proof mainly follows Xiao (2022), but all results are carefully extended for the robust case. In particular, from combining Lemma 1 with the cohesive bond lemma (Appendix C.4), we get directly that $\rho^*_{\mathcal{U}} - \rho^{\pi_{k+1}}_{\mathcal{U}} \leq C_{PL} \max_{\pi' \in \Pi} \langle \nabla \rho^{\pi_{k+1}}_{\mathcal{U}}, \pi' - \pi_{k+1} \rangle \leq 2 C_{PL} L \|\pi_{k+1} - \pi_k\| \mathbf{diam}(\Pi)$, that provides us the lower bound on $\|\pi_{k+1} - \pi_k\|$. In contrast, sufficient increase lemma (Appendix C.2) provides and upper bound on $\|\pi_{k+1} - \pi_k\|$: $\rho^{\pi_{k+1}}_{\mathcal{U}} - \rho^{\pi_k}_{\mathcal{U}} \geq \frac{L}{2} \|\pi_{k+1} - \pi_k\|^2, \forall k$. Combining these upper and lower bounds, we get the following sub-optimality recursion: $a^2_{k+1} + a_{k+1} \leq a_k$, for $a_k := \frac{\rho^*_{\mathcal{U}} - \rho^{\pi_k}_{\mathcal{U}}}{8 L C^2_{PL} \mathbf{diam}(\Pi)^2}$ (Appendix C.5), which in turn implies the convergence. Indeed, note that, the sub-optimality recursion $a_{k+1} - a_k \leq -a^2_{k+1}$ corresponds to an ordinary differential equation $\frac{da}{dk} \leq -a^2$, whose solution is $a(k) \leq \frac{1}{k + \frac{1}{a(0)}} \leq \frac{1}{k}$. This relation intuitively indicates an $O(\frac{1}{\epsilon})$ iteration complexity for achieving an ϵ-optimal solution. □

3 Discussion

We establish global convergence for robust MDPs with L-smooth robust return with respect to a policy, and proved its iteration complexity $O(SA\epsilon^{-1})$, which is much faster than existing state-of-the-art complexity $O(S^4 A^2 \epsilon^{-4})$ for robust MDPs (Wang et al., 2022b). Moreover, our proof trivially yields a simpler and more intuitive proof for non-robust MDPs by taking a single environment uncertainty set.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

REFERENCES

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift, 2020.

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

Bahram Behzadian, Marek Petrik, and Chin Pang Ho. Fast algorithms for l_∞-constrained s-rectangular robust MDPs. *Advances in Neural Information Processing Systems*, 34:25982–25992, 2021.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods, 2019. URL `https://arxiv.org/abs/1906.01786`.

Uri Gadot, Esther Derman, Navdeep Kumar, Maxence Mohamed Elfatihi, Kfir Levy, and Shie Mannor. Solving non-rectangular reward-robust mdps via frequency regularization, 2023.

Vineet Goyal and Julien Grand-Clément. Robust markov decision process: Beyond rectangularity, 2018. URL `https://arxiv.org/abs/1811.00215`.

Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Partial policy iteration for l1-robust markov decision processes, 2020. URL `https://arxiv.org/abs/2006.09484`.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, May 2005a. ISSN 1526-5471. doi: 10.1287/moor.1040.0129. URL `http://dx.doi.org/10.1287/MOOR.1040.0129`.

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

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. In S. Solla, T. Leen, and K. Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL `https://proceedings.neurips.cc/paper/1999/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf`.

Navdeep Kumar, Kfir Levy, Kaixin Wang, and Shie Mannor. Efficient policy iteration for robust markov decision processes via regularization, 2022. URL `https://arxiv.org/abs/2205.14327`.

Navdeep Kumar, Esther Derman, Matthieu Geist, Kfir Levy, and Shie Mannor. Policy gradient for s-rectangular robust markov decision processes, 2023. URL `https://arxiv.org/abs/2301.13589`.

Yan Li, Tuo Zhao, and Guanghui Lan. First-order policy optimization for robust markov decision process, 2022. URL `https://arxiv.org/abs/2209.10579`.

Michael Lutter, Shie Mannor, Jan Peters, Dieter Fox, and Animesh Garg. Robust value iteration for continuous control tasks. *Robotics: Science and Systems*, 2021.

Shie Mannor, Duncan Simester, Peng Sun, and John N Tsitsiklis. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, 2007.

Shie Mannor, Ofir Mebel, and Huan Xu. Robust mdps with k-rectangular uncertainty. *Math. Oper. Res.*, 41(4):1484–1509, nov 2016. ISSN 0364-765X.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6820–6829. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/mei20b.html`.

Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70: 583–601, 02 2002. doi: 10.1111/1468-0262.00296.

Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Oper. Res.*, 53:780–798, 2005.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pp. 2817–2826. PMLR, 2017.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization, 2015. URL https://arxiv.org/abs/1502.05477.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

Richard S Sutton, David A McAllester, Satinder P Singh, Yishay Mansour, et al. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 99, pp. 1057–1063. Citeseer, 1999.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 2000. URL https://proceedings.neurips.cc/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf.

Kaixin Wang, Navdeep Kumar, Kuangqi Zhou, Bryan Hooi, Jiashi Feng, and Shie Mannor. The geometry of robust value functions. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22727–22751. PMLR, 17–23 Jul 2022a. URL https://proceedings.mlr.press/v162/wang22k.html.

Qiuhao Wang, Chin Pang Ho, and Marek Petrik. On the convergence of policy gradient in robust mdps, 2022b. URL https://arxiv.org/abs/2212.10439.

Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty, 2021. URL https://arxiv.org/abs/2109.14523.

Yue Wang and Shaofeng Zou. Policy gradient method for robust reinforcement learning, 2022a.

Yue Wang and Shaofeng Zou. Policy gradient method for robust reinforcement learning. *International Conference on Machine Learning*, 162:23484–23526, 2022b.

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

Berç Rustem Wolfram Wiesemann, Daniel Kuhn. Robust markov decision processes. *Mathematics of Operations Research 38(1):153-183*, 2012.

Lin Xiao. On the convergence rates of policy gradient methods, 2022. URL https://arxiv.org/abs/2201.07443.

Huan Xu and Shie Mannor. Robustness and generalization, 2010. URL https://arxiv.org/abs/1005.2243.

A Appendix

Related Work

Non-Robust MDPs. Policy gradient is derived in Sutton et al. (2000) for non-robust MDPs which is widely used in practice with many variants (Schulman et al., 2015; Konda & Tsitsiklis, 1999; Sutton & Barto, 2018). Recently, there have been global convergences guarantees results Agarwal et al. (2020); Bhandari & Russo (2019) with an iteration complexity $O(1/\epsilon)$ for finding ϵ-optimal policy Xiao (2022).

(s,a)**-rectangular R-Contamination Robust MDPs**. The paper Wang & Zou (2022a) derives policy gradient for R-rectangular robust MDPs complexity $O(S^2 A\log(\frac{1}{\epsilon}))$ similar to non-robust MDPs. Further, it establishes global convergence policy gradient with an iteration complexity $O(1/\epsilon^3)$ for finding ϵ-optimal policy assuming oracle policy gradient.

General (s,a)**-rectangular Robust MDPs** The paper Li et al. (2022) establishes global convergence for robust mirror policy decent for (s,a)-rectangular robust MDPs in general with an iteration complexity $O(1/\epsilon)$ and $O(log(1/\epsilon))$ for finding ϵ-optimal policy, with two increasing-stepsize schemes. However, it assumes the oracle access to policy gradient.

General Robust MDPs The paper Wang et al. (2022b) establishes global convergence for Double-Loop Robust Policy Gradient for general robust MDPs with an iteration complexity $O(1/\epsilon^4)$ for finding ϵ-optimal policy, assuming the oracle access to policy gradient. Solving the policy gradient upto ϵ tolerance via value methods that takes (s,a)-rectangular and s-rectangular case with complexity of $O(S^4 A\log(1/\epsilon))$ and $O(S^4 A^3\log(1/\epsilon))$ respectively using convex optimizations tools. Our techniques are completely different than this work.

B Smoothness

B.1 Background

We go over the some definitions and background results helpful for our discussion. Let value function and Q-value function be defined as follows:

$$v^\pi_{(P,R)} := (I-\gamma P^\pi)^{-1}, \qquad Q^\pi_{(P,R)}(s,a) := R(s,a) + \gamma\sum_{s'} P(s'|s,a)v^\pi_{(P,R)}(s').$$

Following Kumar et al. (2023), let the occupation measure for policy π with initial vector $u\in\mathbb{R}^{\mathcal{S}}$ be defined by

$$d^\pi_{P,u} := u^T(I-\gamma P^\pi)^{-1}.$$

Let $P^\pi_{\mathcal{U}}, R^\pi_{\mathcal{U}}$ be the worst kernel and reward function under policy π, that is:

$$(P^\pi_{\mathcal{U}}, R^\pi_{\mathcal{U}}) \in \arg\min_{(P,R)\in\mathcal{U}} \rho^\pi_{(P,R)}.$$

Similarly to Kumar et al. (2023), we define the robust values as follows

$$d^\pi_{\mathcal{U}} := d^\pi_{P^\pi_{\mathcal{U}}}, \qquad v^\pi_{\mathcal{U}} = v^\pi_{(P^\pi_{\mathcal{U}}, R^\pi_{\mathcal{U}})}, \qquad Q^\pi_{\mathcal{U}} = Q^\pi_{(P^\pi_{\mathcal{U}}, R^\pi_{\mathcal{U}})}.$$

Let $\mathcal{R}^{\mathtt{s}}_p, \mathcal{P}^{\mathtt{s}}_p$ denote $\mathtt{s}$-rectangular L_p-bounded reward noise uncertainty set and s-rectangular L_p-bounded kernel noise uncertainty set respectively, as they are decomposable over states, that is,

$$\mathcal{R}^{\mathtt{s}}_p = \otimes_{s\in\mathcal{S}} R_s, \quad \text{and} \quad \mathcal{P}^{\mathtt{s}}_p = \otimes_{s\in\mathcal{S}} P_s,$$

where $\|R_s\|_p \leq \alpha_s, \|P_s\|_p \leq \beta_s$ and $\sum_{s'} P_s(s'|a) = 0 \forall s,a$ (see Wolfram Wiesemann (2012)). Similarly, $\mathcal{R}^{\mathtt{sa}}, \mathcal{P}^{\mathtt{sa}}$ denotes $\mathtt{sa}$-rectangular reward uncertainty set and kernel respectively, as they are decomposable over states-actions, that is,

$$\mathcal{R}^{\mathtt{sa}}_p = \otimes_{s\in\mathcal{S},a\in\mathcal{A}} R_{sa}, \quad \text{and} \quad \mathcal{P}^{\mathtt{sa}}_p = \otimes_{s\in\mathcal{S},a\in\mathcal{A}} P_{sa},$$

where $\|R_{sa}\|_p \leq \alpha_{sa}, \|P_{sa}\|_p \leq \beta_{sa}$ and $\sum_{s'} P_{sa}(s') = 0 \forall s,a$ (see Iyengar (2005b); Nilim & Ghaoui (2005)).

Let q be the Holder's conjugate of $p\in[1,\infty)$, that satisfying,

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Proposition 1. *The non-robust return $\rho^\pi_{P,R}$ is $\frac{A}{(1-\gamma)^2}$-smooth for $\|R\|_\infty \leq 1$.*

Proof. See Agarwal et al. (2020). □

B.2 SMOOTHNESS RESULTS

Now we are ready to state our first smoothness result of this section.

Proposition 2. *For the non-rectangular reward uncertainty set $\mathcal{U} = \{P_0\} \otimes \{R_0 + \mathcal{R}_p\}$, the robust return $\rho^\pi_\mathcal{U}$ is $2(q-1)\frac{(SA)^{\frac{q+1}{q}}A^2}{(1-\gamma)^4} + 2\gamma\frac{(SA)^{\frac{1}{q}}A}{(1-\gamma)^3}$-smooth, where $\mathcal{R}_p := \{R \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \mid \|R\|_p \leq \alpha\}$.*

Proof. Proved in Gadot et al. (2023). □

Lemma 2. *$\rho^\pi_\mathcal{U}$ is $\frac{\|R\|_\infty + \|\alpha\|_\infty}{(1-\gamma)^3}$-smooth for the uncertainty set $\mathcal{U} = \{P_0\} \times \{R_0 + \mathcal{R}^{sa}_p\}$, where $\|\alpha\|_\infty = \max_{s,a}\alpha_{sa}$.*

Proof. From Kumar et al. (2022), we have

$$\rho^\pi_\mathcal{U} = \rho^\pi_{(P_0,R_0)} - \sum_{s,a} d^\pi_{P_0}(s)\pi(a|s)\alpha_{sa}. \tag{2}$$

Now we know $\langle d^\pi_P, R^\pi\rangle$ is $\frac{\|R\|_\infty}{(1-\gamma)^3}$-smooth Agarwal et al. (2020). Hence the robust return $\rho^\pi_\mathcal{U}$ is $\frac{\|R\|_\infty + \|\alpha\|_\infty}{(1-\gamma)^3}$ - smooth. □

Lemma 3. *$\rho^\pi_\mathcal{U}$ is smooth for the uncertainty set $\mathcal{U} = \{P_0\} \times \{R_0 + \mathcal{R}^s_p\}$ for $p \in (1, \infty)$.*

Proof. From Kumar et al. (2022), we have

$$\rho^\pi_\mathcal{U} = \rho^\pi_{(P_0,R_0)} - \sum_{s} d^\pi_{P_0}(s)\|\pi_s\|_q\alpha_s. \tag{3}$$

It is clear that $\|\pi_s\|_q$ is smooth in π for all $p \in (1, \infty)$. Furthermore, we know $d^\pi_{P_0}(s)$ is also smooth, as $d^\pi_{P_0}(s) = \rho^\pi_{P_0,R}$ where $R(s', a') = \mathbf{1}((s', a') = (s, a))$, and we know $\rho^\pi_{P,R}$ is smooth for all P and all R Agarwal et al. (2020). This establishes the smoothness of the robust return in our case, as it is algebraic function of smooth functions. □

We conclude that there are various setting under which the robust return is smooth, the our global convergence analysis for policy gradient methods applies.

C GLOBAL CONVERGENCE

C.1 ASSUMPTIONS

Assumption 1. *We assume the uncertainty set $\mathcal{U}$ is convex and compact.*

The above assumption is very mild that is satisfied in most of the settings.

Assumption 2. *[**Smoothness**] The function $\rho^\pi_\mathcal{U}$ is L-smooth function, that is*

$$\Big|\ \rho^{\pi'}_\mathcal{U} - \rho^\pi_\mathcal{U} - \langle\nabla\rho^\pi_\mathcal{U}, \pi' - \pi\rangle\ \Big| \leq \frac{L}{2}\|\pi' - \pi\|^2, \qquad \forall\pi', \pi \in \Pi. \tag{4}$$

The assumption doesn't hold for general uncertainty set, however it may hold for many useful uncertainty sets.

Policy udate is done via projected gradient ascent as:

$$\pi_{k+1} := \mathbf{proj}_\Pi(\pi_k + \frac{1}{L}\nabla_\pi\rho^{\pi_k}_\mathcal{U}), \tag{5}$$

where $\frac{1}{L}$ is the learning rate and $\mathbf{proj}_\Pi$ denotes the orthogonal projection onto set Π.

Lemma 4. *(Convex Projection Lemma) For any convex set $\mathbb{X} \subseteq \mathbb{R}^d$, any point $a \in \mathbb{X}$, and any update direction $u \in \mathbb{R}^d$, let $b = proj_\mathbb{X}\ \Big(\ a + u\ \Big)$ be the projection of $a + u$ onto $\mathbb{X}$, then we have*

1. $\langle u, b-a\rangle \geq \|b-a\|_2^2$.
2. $\langle c-b, u-(b-a)\rangle \leq 0, \qquad \forall c \in \mathcal{X}$.

Proof. Follows trivially from the geometry (see figure 1), and the fact that the hyperplane separates a convex set from a point not in the set. □

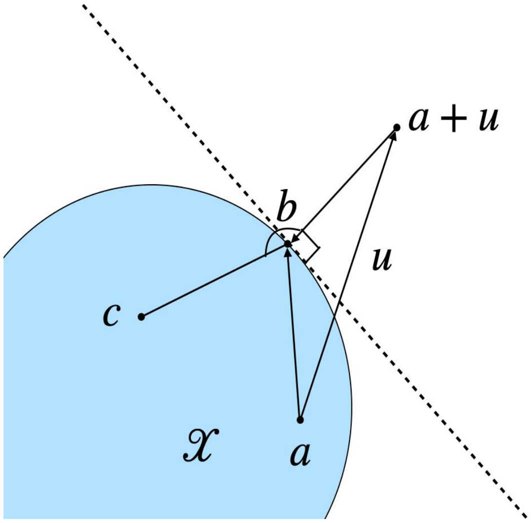

Figure 1: Convex Projection

Remark 1. *For simplicity of the notation, we use ρ_k for $\rho_{\mathcal{U}}^{\pi_k}$.*

C.2 Monotone improvement of the robust return

Now present 'Sufficient Increase Lemma' ensures monotone improvement of the robust return, that uses only convexity of the projection set Π, smoothness of the robust return.

Lemma 5. ***[Sufficient Increase Lemma]** Gradient ascent ensures the monotone improvement in the robust return. Precisely,*

$$\rho_{k+1} - \rho_k \geq \frac{L}{2}\|\pi_{k+1} - \pi_k\|^2, \qquad \forall k.$$

Proof. From smoothness of the robust return, we have

$$\begin{aligned} \rho_{k+1} &\geq \rho_k + \langle \nabla \rho_k, \pi_{k+1} - \pi_k\rangle - \frac{L}{2}\|\pi_{k+1} - \pi_k\|^2, && (6)\\ &= \rho_k + L\langle \frac{1}{L}\nabla \rho_k, \pi_{k+1} - \pi_k\rangle - \frac{L}{2}\|\pi_{k+1} - \pi_k\|^2, && (7)\\ &\geq \rho_k + L\|\pi_{k+1} - \pi_k\|^2 - \frac{L}{2}\|\pi_{k+1} - \pi_k\|^2, \qquad \text{(from Lemma 4).} && (8) \end{aligned}$$

This ends the proof. □

Note that the above sufficient increase lemma does not require the robust return to be concave. Further, the above lemma is enough to ensure iterates $\{\rho_{\mathcal{U}}^{\pi_k}\}$ converge to some some value $\hat{\rho}$, as the iterates forms monotonically increasing sequence. However, it doesn't imply the $\hat{\rho}$ is global maxima or local maxima for that matter. This just implies, the iterates $\rho_{\mathcal{U}}^{\pi_k}$ keeps on increasing until the gradient $G(\pi_k)$ doesn't diminish to zero.

C.3 Gradient Domination

Hence, for the global optimality, we need second part, to ensure that the norm of the gradient vanishes only when the sub-optimality does.

Lemma 6 (**Gradient Domination lemma**)**.** *For any policy $\pi \in \Pi$, its sub-optimality is bounded by its policy gradient as*

$$\rho^*_{\mathcal{U}} - \rho^\pi_{\mathcal{U}} \le C_{\text{PL}} \max_{\pi' \in \Pi} \langle \pi' - \pi, \nabla \rho^\pi_{\mathcal{U}} \rangle,$$

where $C_{\text{PL}} := \max_{(\pi,s)\in\Pi\times\mathcal{S}} \frac{d^{{\pi^*}_{P^\pi_{\mathcal{U}}}}_{P^\pi_{\mathcal{U}}}(s)}{d^\pi_{P^\pi_{\mathcal{U}}}(s)}$.

Proof. Recall, $(P^\pi_{\mathcal{U}}, R^\pi_{\mathcal{U}}) \in \arg\min_{(P,R)\in\mathcal{U}} \rho^\pi_{(P,R)}$ is the worst parameters w.r.t. policy π. And $\pi^*_{(P,R)} \in \arg\max_\pi \rho^\pi_{(P,R)}$ be the best policy for dynamics (P, R). Then we have

$$\begin{aligned}
\rho^*_{\mathcal{U}} - \rho^\pi_{\mathcal{U}} &= \max_{\pi'} \min_{(P,R)\in\mathcal{U}} \rho^{\pi'}_{(P,R)} - \rho^\pi_{(P^\pi_{\mathcal{U}},R^\pi_{\mathcal{U}})}, \qquad \text{(by definition)}, && (9)\\
&\le \max_{\pi'} \rho^{\pi'}_{(P^\pi_{\mathcal{U}},R^\pi_{\mathcal{U}})} - \rho^\pi_{(P^\pi_{\mathcal{U}},R^\pi_{\mathcal{U}})}, \qquad \text{(by definition of min-operator)}, && (10)\\
&\text{(Using non-robust Performance Difference Lemma Agarwal et al. (2020), we get)} && (11)\\
&= \sum_s d^{{\pi^*}_{P^\pi_{\mathcal{U}}}}_{P^\pi_{\mathcal{U}}}(s) \sum_a \big(\, \pi^*_{(P^\pi_{\mathcal{U}},R^\pi_{\mathcal{U}})}(a|s) - \pi(a|s) \,\big)\, Q^\pi_{(P^\pi_{\mathcal{U}},R^\pi_{\mathcal{U}})}(s,a), && (12)\\
&\le \sum_s d^{{\pi^*}_{P^\pi_{\mathcal{U}}}}_{P^\pi_{\mathcal{U}}}(s) \underbrace{\max_{\pi'_s} \sum_a \big(\, \pi'_s(a) - \pi(a|s) \,\big)\, Q^\pi_{\mathcal{U}}(s,a)}_{\ge 0} && (13)\\
&= \sum_s \frac{d^{{\pi^*}_{P^\pi_{\mathcal{U}}}}_{P^\pi_{\mathcal{U}}}(s)}{d^\pi_{\mathcal{U}}(s)} d^\pi_{\mathcal{U}}(s) \underbrace{\max_{\pi'_s} \sum_a \big(\, \pi'_s(a) - \pi(a|s) \,\big)\, Q^\pi_{\mathcal{U}}(s,a)}_{\ge 0} && (14)\\
&\le \Big(\, \max_s \frac{d^{{\pi^*}_{P^\pi_{\mathcal{U}}}}_{P^\pi_{\mathcal{U}}}(s)}{d^\pi_{\mathcal{U}}(s)} \,\Big) \max_{\pi'} \sum_s d^\pi_{\mathcal{U}}(s) \sum_a \big(\, \pi'(a|s) - \pi(a|s) \,\big)\, Q^\pi_{\mathcal{U}}(s,a) && (15)\\
&= \Big(\, \max_s \frac{d^{{\pi^*}_{P^\pi_{\mathcal{U}}}}_{P^\pi_{\mathcal{U}}}(s)}{d^\pi_{\mathcal{U}}(s)} \,\Big) \max_{\pi'} \Big\langle\, \pi' - \pi, \nabla \rho^\pi_{\mathcal{U}} \,\Big\rangle\, . && (16)
\end{aligned}$$

The last equlality comes from the Envelope theorem Milgrom & Segal (2002) and policy gradient theorem Sutton et al. (1999),

$$\frac{\partial \rho^\pi_{\mathcal{U}}}{\partial \pi(a|s)} = \sum_s d^\pi_{P^\pi_{\mathcal{U}}}(s) Q^\pi_{(P^\pi_{\mathcal{U}},R^\pi_{\mathcal{U}})}(s,a).$$

□

C.4 Cohesive bond

Now, we have both the parts: One that lower bounds the gradient and the other that upper bounds it. However, they are not exactly in very compatible forms, hence we require the result below that acts a cohesive bond between the two.

Lemma 7. *For all π_k, we have*

$$\langle \nabla \rho_{k+1}, \pi' - \pi_{k+1} \rangle \le 2L \|\pi_{k+1} - \pi_k\| \mathbf{diam}(\Pi),$$

where $\mathbf{diam}(C) := \max_{x,y\in C} \|x - y\|$ *is the diameter of* C.

Proof. For all $x, y \in C$, we have:

$$
\begin{aligned}
&\langle \nabla \rho_{k+1}, \pi' - \pi_{k+1} \rangle \\
&= \langle \nabla \rho_{k+1} - \nabla \rho_k + \nabla \rho_k, \pi' - \pi_{k+1} \rangle, \qquad \text{[Subtract \& add } \nabla \rho_k\text{]} \\
&= \langle \nabla \rho_{k+1} - \nabla \rho_k, \pi' - \pi_{k+1} \rangle + \langle \nabla \rho_k, \pi' - \pi_{k+1} \rangle, \qquad \text{[Linearity of scalar product]} \\
&\leq \| \nabla \rho_{k+1} - \nabla \rho_k \| \| \pi' - \pi_{k+1} \| + \langle \nabla \rho_k, \pi' - \pi_{k+1} \rangle, \text{Cauchy-Schwartz inequality]} \\
&\leq L \| \pi_{k+1} - \pi_k \| \| \pi' - \pi_{k+1} \| + \langle \nabla \rho_k, \pi' - \pi_{k+1} \rangle, \qquad \text{[Smoothness of robust return]} \\
&= L \| \pi_{k+1} - \pi_k \| \| \pi' - \pi_{k+1} \| + L \langle \frac{1}{L} \nabla \rho_k - (\pi_{k+1} - \pi_k) + (\pi_{k+1} - \pi_k), \pi' - \pi_{k+1} \rangle \\
&\leq 2L \| \pi_{k+1} - \pi_k \| \| \pi' - \pi_{k+1} \|, L \langle \frac{1}{L} \nabla \rho_k - (\pi_{k+1} - \pi_k), \pi' - \pi_{k+1} \rangle \qquad \text{[Cauchy-Schwartz inequality]} \\
&\leq 2L \| \pi_{k+1} - \pi_k \| \| \pi' - \pi_{k+1} \|, \qquad \text{[From Lemma 4]} \\
&\leq 2L \| \pi_{k+1} - \pi_k \| \mathbf{diam}(\Pi).
\end{aligned}
$$

□

C.5 Proof of global convergence

Now we are ready to prove our core result that is, the sub-optimality recursion.

Lemma. *Take $\eta = \frac{1}{L}$ as a learning rate. Then, the scaled sub-optimality $a_k = \frac{\rho_{\mathcal{U}}^* - \rho_{\mathcal{U}}^{\pi_k}}{8LC_{PL}^2 \mathbf{diam}(\Pi)^2}$ follows the recursion*

$$a_{k+1}^2 + a_{k+1} - a_k \leq 0.$$

Proof. From the PL condition proved in Lemma 1, we have

$$
\begin{aligned}
\rho_{\mathcal{U}}^* - \rho_{\mathcal{U}}^{\pi_{k+1}} &\leq C_{PL} \max_{\pi'} \langle \pi' - \pi_{k+1}, \nabla \rho_{\mathcal{U}}^{\pi_{k+1}} \rangle && (17) \\
&\leq 2L \| \pi_{k+1} - \pi_k \| \mathbf{diam}(\Pi), \qquad \text{(from Lemma 7)} && (18) \\
&\leq C_{\text{PL}} \cdot 2 \sqrt{2L(\rho_{\pi}^{\pi_{k+1}} - \rho_{\mathcal{U}}^{\pi_k})} \cdot \mathbf{diam}(\Pi), \qquad \text{(from Lemma 5)} && (19)
\end{aligned}
$$

Squaring both sides and adding subtracting ρ^* in RHS, we get

$$\Big(\rho_{\mathcal{U}}^* - \rho_{\mathcal{U}}^{\pi_{k+1}} \Big)^2 \leq 8C_{\text{PL}}^2 L \mathbf{diam}(\Pi)^2 \Big((\rho_{\mathcal{U}}^* - \rho_{\mathcal{U}}^{\pi_k}) + (\rho_{\mathcal{U}}^* - \rho_{\pi}^{\pi_{k+1}}) \Big)$$

Setting $a_k := \frac{\rho_{\mathcal{U}}^* - \rho_{\mathcal{U}}^{\pi_k}}{8LC_{\text{PL}}^2 \mathbf{diam}(\Pi)^2}$, the sequence $(a_k)_{k \in \mathbb{N}}$ satisfies the recursion $a_{k+1}^2 \leq a_k - a_{k+1}$. □

The sub-optimality recursion derived in the theorem above, illustrates how the sub-optimality at time $k+1$ depends at the sub-optimality at time k. Moreover, the sub-optimality recursion has the quadratic form and $a_k \geq 0$, hence its solution is given as

$$a_{k+1} \leq \sqrt{\frac{1}{4} + a_k} - \frac{1}{2}.$$

As a sanity check, we observe that $\sqrt{\frac{1}{4} + a} - \frac{1}{2} \leq a$ for all $a \geq 0$, implying that $(a_k)_{k \in \mathbb{N}}$ is monotonically decreasing. Further, 0 is the only non-negative fixed point of the $\sqrt{\frac{1}{4} + a} - \frac{1}{2} = a$ implying that $(a_k)_{k \in \mathbb{N}}$ monotonically decreases to 0.

Now, we investigate the convergence rate for a_k. Observe that if $a_0, a_k \gg 1$, then $a_{k+1} \approx \sqrt{a_k}$ and $a_k \approx (a_0)^{\frac{1}{2^k}}$. That is, the convergence rate is super-exponential! Yet, in most cases, $8LC_{\text{PL}}^2 \mathbf{diam}(\Pi)^2 \gg 1$ and $\rho_{\mathcal{U}}^* - \rho_{\mathcal{U}}^{\pi_0} = 8LC_{\text{PL}}^2 \mathbf{diam}(\Pi)^2 a_0 = O(1)$ is bounded so we are more interested in the case where $a_0 \ll 1$. In fact, in an MDP with a reward smaller than 1, we do have $\rho_{\mathcal{U}}^{\pi} = O(1)$.

In this regime, the sub-optimality recursion $a_{k+1} - a_k \leq -a_{k+1}^2$ suggests the ordinary differential equation $\frac{da}{dk} \leq -a^2$ whose solution is $a(k) \leq \frac{1}{k + \frac{1}{a(0)}} \leq \frac{1}{k}$. This intuitively indicates an $O(\frac{1}{\epsilon})$ iteration complexity for achieving an ϵ-optimal solution, which we state below formally.

Corollary (**Global optimality**). *For all iterations $k \geq 1$, it holds that:*

$$\rho^*_{\mathcal{U}} - \rho^{\pi_k}_{\mathcal{U}} \leq \max\left(\frac{8LC^2_{\text{PL}}\mathbf{diam}(\Pi)^2}{k}, 2^{-\frac{k}{2}}\right)(\rho^*_{\mathcal{U}} - \rho^{\pi_0}_{\mathcal{U}}).$$

Proof. The sub-optimality recursion yields the desired result which follows directly from Xiao (2022). □

It can be seen that he exponential part is always lower than the other term. Hence, we can ignore the exponential term. Further, the diameter of the policy class Π, can be upper bounded as

$$\mathbf{diam}(\Pi)^2 = \max_{\pi,\pi}\sum_s \|\pi'_s - \pi_s\|^2_2 \leq \max_{\pi',\pi}\sum_s \|\pi'_s - \pi_s\|^2_1 \leq 4S.$$

This yields the desired result.

In addition, it can be noted that our result reduces to non-robust case, when we take uncertainty set $\mathcal{U}$ to be a singleton set. In this case, our proof is simplified and more-readable version of Agarwal et al. (2020); Xiao (2022).