Policy Gradient Method For Robust Reinforcement Learning

Yue Wang, Shaofeng Zou

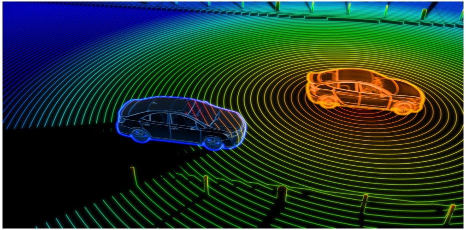
Department of Electrical Engineering
University at Buffalo

2022 International Conference on Machine Learning
Jul 2022

What is Reinforcement Learning (RL)

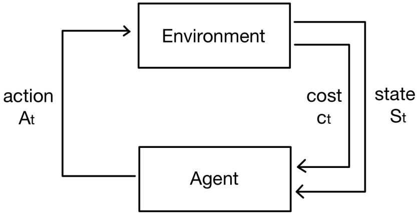
Learn what to do/ how to make decisions

(a) Alpha GO



(b) Autonomous Driving

Interaction Between Agent and Environment



Markov decision process (MDP): $(\mathcal{S}, \mathcal{A}, \mathsf{P}, c, \gamma)$

$\mathcal{S}$: state space

$\mathcal{A}$: action space

P: transition kernel

c: cost function

γ: discount factor

Motivation for Robust RL

In practice, the training environment may be different from the test environment, resulting in a model mismatch, e.g.,

- modeling error between simulator and real-world applications
- model deviation due to non-stationarity of the environment
- unexpected perturbation and potential adversarial attacks.

Goal: find a policy performs well under model mismatch

Robust RL under Model Uncertainty

Robust MDP: $(\mathcal{S}, \mathcal{A}, \mathcal{P}, c, \gamma)$

- $\mathcal{P}$: uncertainty set of transition kernels
- Transition kernel at each time step comes from $\mathcal{P}$, and may be time-varying: $\kappa = (\mathrm{P}_0, \mathrm{P}_1, ...) \in \bigotimes_{t \geq 0} \mathcal{P}$

Pessimistic approach in face of uncertainty:

- (robust value function) $V^\pi(s) = \max_{\kappa \in \otimes_{t\geq 0} \mathcal{P}} \mathbb{E}_\kappa \left[\sum_{t=0}^{\infty} \gamma^t c(S_t, A_t) | S_0 = s, \pi\right]$
- Aims to provide a worst-case performance guarantee

Goal: Optimize the **worst-case** performance $\min_\pi J_\rho(\pi) \triangleq \min_\pi \mathbb{E}_\rho[V^\pi(S)]$

Related Works

Adversarial Robust RL (Vinitsky et al., 2020; Pinto et al., 2017; Abdullah et al., 2019; Hou et al., 2020; Rajeswaran et al., 2017; Huang et al., 2017; Kos and Song, 2017; Pattanaik et al., 2018; Mandlekar et al., 2017), etc. *Empirical success but lack of theoretical understanding*

Model-Based Robust MDP (Iyengar, 2005; Nilim and El Ghaoui, 2004; Bagnell et al., 2001; Satia and Lave Jr, 1973; Wiesemann et al., 2013; Tamar et al., 2014). *Assume knowledge of uncertainty set and solve using dynamic programming*

Model-Free Value-based Method (Roy et al., 2017; Badrinath and Kalathil, 2021). *Not well-justified relaxation on uncertainty sets, strict assumptions on discounted factor;* (Wang and Zou, 2021). *Value-based method, costly when* $\mathcal{S}, \mathcal{A}$ *are large*

Main Contributions

We develop the first direct policy search method with global optimality for model-free robust RL problems, and further characterize its sample complexity

Major Challenges and Contributions

Robust value function V^π may not be differentiable and non-convex

$V^\pi(s) = \max_{\kappa \in \otimes_{t \geq 0} \mathcal{P}} \mathbb{E}_\kappa \left[\sum_{t=0}^{\infty} \gamma^t c(S_t, A_t) | S_0 = s, \pi\right]$ is non-differentiable because of the max operator

- Generalize the vanilla policy gradient to the robust policy sub-gradient method, which shows global optimality
- Develop a smoothed robust policy gradient method with global optimality and $\mathcal{O}(\epsilon^{-3})$ sample complexity
- Show a convex-like proposition (PL-condition) and global optimality

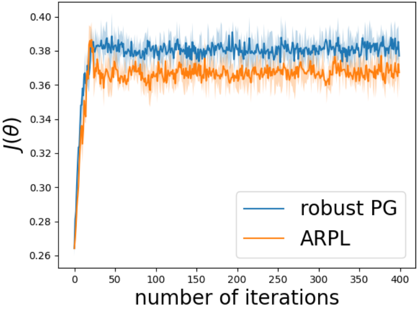
Major Challenges and Contributions

In model-free setting, robust value functions measure the worst-case performance and are impossible to estimate using Monte Carlo method

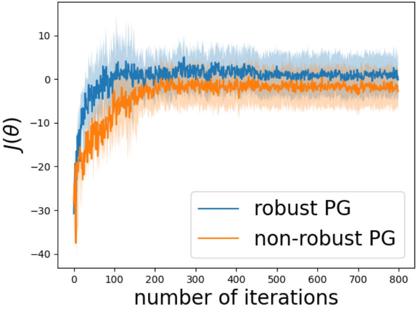
- Propose a robust TD algorithm (which can be applied together with function approximation) to estimate the value functions, and further develop a robust actor-critic algorithm

Numerical Experiments

Experiments show that our methods are more robust to the model mismatch than non-robust methods and some adversarial methods (e.g., ARPL Mandlekar et al. (2017))



(a) Compression on Garnet problem



(b) Compression on Taxi problem

We trained algorithms under an unperturbed MDP, and evaluate their performance under the worst-case transition kernel.

Conclusion

We developed a direct policy search method with provable global optimality for robust RL problems.
Our method is robust to model uncertainty and can be applied with function approximation.

Thanks for listening!

Reference I

Abdullah, M. A., Ren, H., Ammar, H. B., Milenkovic, V., Luo, R., Zhang, M., and Wang, J. (2019). Wasserstein robust reinforcement learning. *arXiv preprint arXiv:1907.13196*.

Badrinath, K. P. and Kalathil, D. (2021). Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *Proc. International Conference on Machine Learning (ICML)*, pages 511–520. PMLR.

Bagnell, J. A., Ng, A. Y., and Schneider, J. G. (2001). Solving uncertain Markov decision processes.

Hou, L., Pang, L., Hong, X., Lan, Y., Ma, Z., and Yin, D. (2020). Robust reinforcement learning with Wasserstein constraint. *arXiv preprint arXiv:2006.00945*.

Huang, S., Papernot, N., Goodfellow, I., Duan, Y., and Abbeel, P. (2017). Adversarial attacks on neural network policies. In *Proc. International Conference on Learning Representations (ICLR)*.

Reference II

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

Kos, J. and Song, D. (2017). Delving into adversarial attacks on deep policies. In *Proc. International Conference on Learning Representations (ICLR)*.

Mandlekar, A., Zhu, Y., Garg, A., Fei-Fei, L., and Savarese, S. (2017). Adversarially robust policy learning: Active construction of physically-plausible perturbations. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3932–3939. IEEE.

Nilim, A. and El Ghaoui, L. (2004). Robustness in Markov decision problems with uncertain transition matrices. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 839–846.

Pattanaik, A., Tang, Z., Liu, S., Bommannan, G., and Chowdhary, G. (2018). Robust deep reinforcement learning with adversarial attacks. In *Proc. International Conference on Autonomous Agents and MultiAgent Systems*, pages 2040–2042.

Reference III

Pinto, L., Davidson, J., Sukthankar, R., and Gupta, A. (2017). Robust adversarial reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, pages 2817–2826. PMLR.

Rajeswaran, A., Ghotra, S., Ravindran, B., and Levine, S. (2017). Epopt: Learning robust neural network policies using model ensembles. In *Proc. International Conference on Learning Representations (ICLR)*.

Roy, A., Xu, H., and Pokutta, S. (2017). Reinforcement learning under model mismatch. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 3046–3055.

Satia, J. K. and Lave Jr, R. E. (1973). Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740.

Tamar, A., Mannor, S., and Xu, H. (2014). Scaling up robust MDPs using function approximation. In *Proc. International Conference on Machine Learning (ICML)*, pages 181–189. PMLR.

Reference IV

Vinitsky, E., Du, Y., Parvate, K., Jang, K., Abbeel, P., and Bayen, A. (2020). Robust reinforcement learning using adversarial populations. *arXiv preprint arXiv:2008.01825*.

Wang, Y. and Zou, S. (2021). Online robust reinforcement learning with model uncertainty. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*.

Wiesemann, W., Kuhn, D., and Rustem, B. (2013). Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183.