Robust Reinforcement Learning under Model Uncertainty

Shaofeng Zou

University at Buffalo, the State University of New York

6th Worshop on Cognition & Control
Jan. 27th, 2023

Outline

Outline

1. Introduction

2. Robust Average-Cost RL
   - Model-based methods
   - Model-free methods

3. Summary

Challenge of Model Mismatch

$\neq$

Training environment $\neq$ test environment

$\Rightarrow$ Model mismatch

$\Rightarrow$ Severe performance degradation

Challenge of Model Mismatch

$\neq$

Training environment $\neq$ test environment

$\Rightarrow$ Model mismatch

$\Rightarrow$ Severe performance degradation

- modeling error between simulator and real-world applications
- non-stationary environment
- unexpected perturbations and potential adversarial attacks

Challenge of Model Mismatch

 $\neq$ 

Training environment $\neq$ test environment

$\Rightarrow$ Model mismatch

$\Rightarrow$ Severe performance degradation

- modeling error between simulator and real-world applications
- non-stationary environment
- unexpected perturbations and potential adversarial attacks

Robust RL:

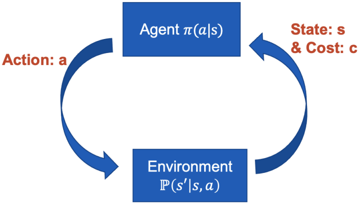
Find good policy that performs well under model mismatch

Markov Decision Processes

- An agent interacts with a stochastic environment: Markov Decision Process (MDP)
- MDP $(\mathcal{S}, \mathcal{A}, \mathsf{P}, c)$

Markov Decision Processes

- An agent interacts with a stochastic environment: Markov Decision Process (MDP)
- MDP $(\mathcal{S}, \mathcal{A}, \mathrm{P}, c)$
  - $\mathcal{S}$: state space
  - $\mathcal{A}$: action space
  - P: transition kernel
  - c: cost function



- A stationary policy $\pi(a|s)$ is a conditional distribution over $\mathcal{A}$

Markov Decision Processes

- Discounted value function for policy π at state s:

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) | S_0 = s, \pi\right]$$

Markov Decision Processes

- Discounted value function for policy π at state s:

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) | S_0 = s, \pi\right]$$

- Goal: find an optimal policy that minimizes value function

$$\min_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) | S_0 = s, \pi\right]$$

Robust MDPs

- 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 $\mathcal{P}$: $\kappa = (\mathsf{P}_0, \mathsf{P}_1, ...) \in \otimes_{t \geq 0} \mathcal{P}$

Robust MDPs

- 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 $\mathcal{P}$: $\kappa = (\mathsf{P}_0, \mathsf{P}_1, ...) \in \otimes_{t \geq 0} \mathcal{P}$
- Pessimistic approach in face of uncertainty:
  - Worst-case overall cost over uncertainty set
  - Robust value function under the discounted setting:

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

  - Goal: Optimize the worst-case performance:

$$\min_{\pi} V^{\pi}_{\mathcal{P},\gamma}(s), \forall s \in \mathcal{S}$$

Related Works: Robust Discounted RL

- Model-based method: $\mathcal{P}$ is known
  - e.g., Bagnell et al. (2001); Nilim and El Ghaoui (2004); Iyengar (2005); Wiesemann et al. (2013); Tamar et al. (2014): robust dynamic programming

Related Works: Robust Discounted RL

- Model-based method: $\mathcal{P}$ is known
  - e.g., Bagnell et al. (2001); Nilim and El Ghaoui (2004); Iyengar (2005); Wiesemann et al. (2013); Tamar et al. (2014): robust dynamic programming
- Model-free method: $\mathcal{P}$ is unknown, samples from nominal transition kernel are available
  - Roy et al. (2017); Panaganti et al. (2022): *ellipsoid-structure* uncertainty set
    - convergence needs discount factor bounded away from 1
  - Liu et al. (2022): KL divergence uncertainty set
    - generative model, tabular setting
  - Wang and Zou (2021, 2022) (our work): R-contamination model
    - asymptotic convergence with sample complexity analysis
    - function approximation
  - Zhou et al. (2021); Yang et al. (2021): offline and tabular

Related Works: Robust Average-Cost RL

$$g_{\mathcal{P}}^{\pi}(s) = \max_{\kappa \in \bigotimes_{t \geq 0} \mathcal{P}} \lim_{n \to \infty} \mathbb{E}_{\kappa} \left[\frac{1}{n} \sum_{t=0}^{n-1} c(S_t, A_t) | S_0 = s, \pi \right]$$

- Model-based method: $\mathcal{P}$ is known
  - Tewari and Bartlett (2007): bounded-interval uncertainty set, limit method
  - Wang et al. (2023b) (our work): general uncertainty set, robust average-cost Bellman equation, limit method and direct method

Related Works: Robust Average-Cost RL

$$g_{\mathcal{P}}^{\pi}(s) = \max_{\kappa \in \bigotimes_{t \geq 0} \mathcal{P}} \lim_{n \to \infty} \mathbb{E}_{\kappa} \left[\frac{1}{n} \sum_{t=0}^{n-1} c(S_t, A_t) | S_0 = s, \pi \right]$$

- Model-based method: $\mathcal{P}$ is known
  - Tewari and Bartlett (2007): bounded-interval uncertainty set, limit method
  - Wang et al. (2023b) (our work): general uncertainty set, robust average-cost Bellman equation, limit method and direct method
- Model-free method: $\mathcal{P}$ is unknown
  - Wang et al. (2023a) (our work): general uncertainty set, robut relative value iteration with convergence guarantee

Heuristic Approaches

- Adversarial state transition perturbation: an adversary perturbs the state transition: Vinitsky et al. (2020); Pinto et al. (2017); Abdullah et al. (2019); Hou et al. (2020); Rajeswaran et al. (2017); Atkeson and Morimoto (2003); Morimoto and Doya (2005)

Heuristic Approaches

- Adversarial state transition perturbation: an adversary perturbs the state transition: Vinitsky et al. (2020); Pinto et al. (2017); Abdullah et al. (2019); Hou et al. (2020); Rajeswaran et al. (2017); Atkeson and Morimoto (2003); Morimoto and Doya (2005)
- Adversarial sample perturbation: an adversary modifies state observations Huang et al. (2017); Kos and Song (2017); Lin et al. (2017); Pattanaik et al. (2018); Mandlekar et al. (2017)

Outline

Robust Average-Cost RL

Recall the robust average-cost:

$$g_{\mathsf{P}}^{\pi}(s) \triangleq \lim_{n\to\infty} \mathbb{E}_{\mathsf{P}}\left[\frac{1}{n}\sum_{t=0}^{n-1} c_t | S_0 = s, \pi\right]$$

$$g_{\mathcal{P}}^{\pi} = \max_{\mathsf{P}\in\mathcal{P}} g_{\mathsf{P}}^{\pi}$$

Goal: Find $\pi^* = \arg\min_{\pi} g_{\mathcal{P}}^{\pi}$

Main Results

- Fundamental understanding of robust average-cost MDPs
  - robust average-cost Bellman equation
- Model-based methods:
  - Limit method
  - Direct method
- Model-free methods: robust TD and robust Q-learning

Robust Average-Cost MDP: Limit Method

Non-robust setting

(Puterman (1994) point-wise convergence) For any fixed P and π, $\lim_{\gamma \to 1}(1-\gamma)V^{\pi}_{\mathsf{P},\gamma} = g^{\pi}_{\mathsf{P}}$

- Under non-robust setting, average-cost can be approximated by discounted value function

Robust Average-Cost MDP: Limit Method

Non-robust setting

(Puterman (1994) point-wise convergence) For any fixed P and π, $\lim_{\gamma\to 1}(1-\gamma)V^{\pi}_{\mathsf{P},\gamma} = g^{\pi}_{\mathsf{P}}$

- Under non-robust setting, average-cost can be approximated by discounted value function
- In robust MDP, does it hold that

$$\lim_{\gamma\to 1}(1-\gamma)V^{\pi}_{\mathcal{P},\gamma} = g^{\pi}_{\mathcal{P}} \quad \textbf{?}$$

Robust Average-Cost MDP: Limit Method

Non-robust setting

(Puterman (1994) point-wise convergence) For any fixed P and π, $\lim_{\gamma\to 1}(1-\gamma)V^{\pi}_{\mathsf{P},\gamma} = g^{\pi}_{\mathsf{P}}$

- Under non-robust setting, average-cost can be approximated by discounted value function
- In robust MDP, does it hold that

$$\lim_{\gamma\to 1}(1-\gamma)V^{\pi}_{\mathcal{P},\gamma} = g^{\pi}_{\mathcal{P}} \quad \textbf{?}$$

- Robust discounted Bellman operator (Nilim and El Ghaoui, 2004; Iyengar, 2005): $\mathbf{T}V = c + \gamma\sum_a \pi(a|s)\sigma_{\mathcal{P}^a_s}(V)$, where $\sigma_{\mathcal{P}^a_s}(V) = \max_{p\in\mathcal{P}^a_s} p^\top V$ is support function
- $\mathbf{T}$ is a γ-contraction and has $V^{\pi}_{\mathcal{P},\gamma}$ as its unique fixed point: $\mathbf{T}V^{\pi}_{\mathcal{P},\gamma} = V^{\pi}_{\mathcal{P},\gamma}$

Tewari and Bartlett (2007): Bounded-interval Uncertainty Set

- Number of possible worst-case transition kernels is finite
  - Proof of this argument relies on structure of bounded-interval
- Then, $\min_P$ and $\lim_\gamma$ are interchangeable
- Not generalizable to general uncertainty sets

Robust Average-Cost MDP: Limit Method

Theorem: uniform convergence

$$\lim_{\gamma \to 1} (1-\gamma) V^{\pi}_{\mathsf{P},\gamma} = g^{\pi}_{\mathsf{P}} \text{ uniformly}$$

Robust Average-Cost MDP: Limit Method

Theorem: uniform convergence

$$\lim_{\gamma \to 1} (1-\gamma) V^{\pi}_{\mathsf{P},\gamma} = g^{\pi}_{\mathsf{P}} \text{ uniformly}$$

- Then $\min_{\mathsf{P}}$ and $\lim_{\gamma \to 1}$ are interchangeable:

$$\begin{aligned} g^{\pi}_{\mathcal{P}} &= \min_{\mathsf{P} \in \mathcal{P}} g^{\pi}_{\mathsf{P}} \\ &= \min_{\mathsf{P} \in \mathcal{P}} \lim_{\gamma \to 1} (1-\gamma) V^{\pi}_{\mathsf{P},\gamma} = \lim_{\gamma \to 1} \min_{\mathsf{P} \in \mathcal{P}} (1-\gamma) V^{\pi}_{\mathsf{P},\gamma} \\ &= \lim_{\gamma \to 1} (1-\gamma) V^{\pi}_{\mathcal{P},\gamma} \end{aligned}$$

Robust Average-Cost MDP: Limit Method

Theorem: uniform convergence

$$\lim_{\gamma\to 1}(1-\gamma)V^{\pi}_{\mathsf{P},\gamma} = g^{\pi}_{\mathsf{P}} \text{ uniformly}$$

- Then $\min_{\mathsf{P}}$ and $\lim_{\gamma\to 1}$ are interchangeable:

$$\begin{aligned} g^{\pi}_{\mathcal{P}} &= \min_{\mathsf{P}\in\mathcal{P}} g^{\pi}_{\mathsf{P}} \\ &= \min_{\mathsf{P}\in\mathcal{P}} \lim_{\gamma\to 1}(1-\gamma)V^{\pi}_{\mathsf{P},\gamma} = \lim_{\gamma\to 1}\min_{\mathsf{P}\in\mathcal{P}}(1-\gamma)V^{\pi}_{\mathsf{P},\gamma} \\ &= \lim_{\gamma\to 1}(1-\gamma)V^{\pi}_{\mathcal{P},\gamma} \end{aligned}$$

- Robust average-cost can be approximated by discounted robust value functions as $\gamma \to 1$

Robust Average-Cost MDP: Limit Method

Basic idea of limit method:

- Set $\gamma_t \to 1$
- Apply one-step robust discounted Bellman operator

Robust value iteration for policy evaluation

INPUT: $\pi, V_0(s) = 0, \forall s, T$
FOR $t = 0, 1, ..., T-1$
$\gamma_t \leftarrow \frac{t+1}{t+2}$
FOR all $s \in \mathcal{S}$
$V_{t+1}(s) \leftarrow \mathbb{E}_\pi[(1-\gamma_t)c(s,A) + \gamma_t \sigma_{\mathcal{P}_s^A}(V_t)]$
OUTPUT: V_T

Robust Average-Cost MDP: Limit Method

Convergence of Robust Value Iteration

RVI algorithm converges to robust average-cost: $\lim_{T\to\infty} V_T \to g_{\mathcal{P}}^{\pi}$

Robust Average-Cost MDP: Limit Method

Convergence of Robust Value Iteration

RVI algorithm converges to robust average-cost: $\lim_{T\to\infty} V_T \to g_{\mathcal{P}}^{\pi}$

- Solves the policy evaluation problem under the robust average-cost setting

Robust Average-Cost MDP: Limit Method

Convergence of Robust Value Iteration

RVI algorithm converges to robust average-cost: $\lim_{T\to\infty} V_T \to g_{\mathcal{P}}^{\pi}$

- Solves the policy evaluation problem under the robust average-cost setting
- Convergence rate: $\|V_T - g_{\mathcal{P}}^{\pi}\| = \mathcal{O}\left(\frac{1}{T}\right)$

Robust Average-Cost MDP: Limit Method

- The limit method also works for optimal control problems

Robust value iteration for optimal control

INPUT: $V_0(s) = 0, \forall s, T$

FOR $t = 0, 1, ..., T-1$

$\gamma_t \leftarrow \frac{t+1}{t+2}$

FOR all $s \in \mathcal{S}$

$$V_{t+1}(s) \leftarrow \min_{a \in \mathcal{A}} \left\{(1-\gamma_t)c(s,a) + \gamma_t \sigma_{\mathcal{P}_s^a}(V_t)\right\}$$

FOR $s \in \mathcal{S}$

$$\pi_t(s) \leftarrow \arg\min_{a \in \mathcal{A}} \left\{(1-\gamma_t)c(s,a) + \gamma_t \sigma_{\mathcal{P}_s^a}(V_t)\right\}$$

OUTPUT: π_T, V_T

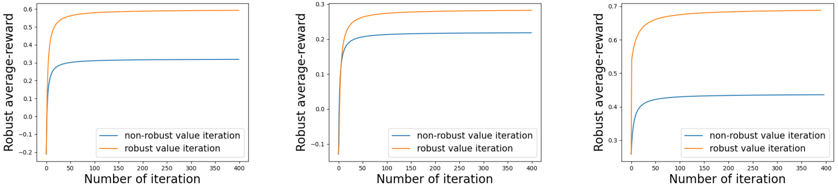
Convergence of robust value iteration

$V_T \to g^*_{\mathcal{P}}$, $\pi_T \to \pi^* = \arg\min_\pi g^\pi_{\mathcal{P}}$

Robust Average-Cost MDP: Limit Method

Non-robust value iteration vs robust value iteration:

- under different uncertainty sets (contamination model, total variation model and KL-divergence model)
- evaluate worst-case performance of obtained policy



RVI is more robust than non-robust value iteration method

Robust Average-Cost MDP: Limit Method

Robust Blackwell optimality

There exists $\delta < 1$, such that for any $\delta < \gamma < 1$, if $\pi^* = \arg\min_\pi V^\pi_{\mathcal{P},\gamma}$ is optimal to robust discounted value function, then π^* is also optimal to robust average-cost $\pi^* \in \arg\min_\pi g^\pi_{\mathcal{P}}$.

Robust Average-Cost MDP: Limit Method

Robust Blackwell optimality

There exists $\delta < 1$, such that for any $\delta < \gamma < 1$, if $\pi^* = \arg\min_\pi V^\pi_{\mathcal{P},\gamma}$ is optimal to robust discounted value function, then π^* is also optimal to robust average-cost $\pi^* \in \arg\min_\pi g^\pi_{\mathcal{P}}$.

- Fundamental relationship between the robust discounted MDPs and robust average-cost MDPs

Robust Average-Cost MDP: Limit Method

Robust Blackwell optimality

There exists $\delta < 1$, such that for any $\delta < \gamma < 1$, if $\pi^* = \arg\min_\pi V^\pi_{\mathcal{P},\gamma}$ is optimal to robust discounted value function, then π^* is also optimal to robust average-cost $\pi^* \in \arg\min_\pi g^\pi_{\mathcal{P}}$.

- Fundamental relationship between the robust discounted MDPs and robust average-cost MDPs
- Analog to Blackwell optimality of non-robust setting

Robust Average-Cost MDP: Limit Method

Robust Blackwell optimality

There exists $\delta < 1$, such that for any $\delta < \gamma < 1$, if $\pi^* = \arg\min_\pi V^\pi_{\mathcal{P},\gamma}$ is optimal to robust discounted value function, then π^* is also optimal to robust average-cost $\pi^* \in \arg\min_\pi g^\pi_{\mathcal{P}}$.

- Fundamental relationship between the robust discounted MDPs and robust average-cost MDPs
- Analog to Blackwell optimality of non-robust setting
- Proofs of non-robust setting and bounded-interval uncertainty set Tewari and Bartlett (2007): for two policies π and ν: $f_{\pi,\nu}(\gamma) \triangleq V^\pi_{\mathsf{P},\gamma} - V^\mu_{\mathsf{P},\gamma}$ is rational function, thus has finite many zeros. This does not hold in robust setting as $V^\pi_{\mathcal{P},\gamma} - V^\mu_{\mathcal{P},\gamma}$ is not rational due to max

Robust Average-Cost MDP: Limit Method

The limit method

- solves robust average-cost MDPs using robust discounted MDPs as intermediate steps
- based on robust discounted MDPs, does not directly study the fundamental structure of robust average-cost MDPs

Robust Average-Cost MDP: Direct Method

Assumption

The Markov chain induced by any $\mathrm{P} \in \mathcal{P}$ and any π is a unichain.

Optimal robust Bellman equation

If (g, V) is a solution to

$$V(s) = \min_a \left\{ c(s,a) - g + \sigma_{\mathcal{P}_s^a}(V) \right\}, \forall s,$$

then $g = g^*_{\mathcal{P}}$. If we further set

$$\pi^*(s) = \arg\min_a \left\{ c(s,a) + \sigma_{\mathcal{P}_s^a}(V) \right\}$$

for any $s \in \mathcal{S}$, then π^* is an optimal robust policy.

- Solving robust average-cost MDPs can be done by solving the robust Bellman equation

$$V(s) = \mathbf{T}(V) = \min_a \left\{ c(s,a) - g + \sigma_{\mathcal{P}_s^a}(V) \right\}$$

How to solve the robust Bellman equation?

Robust Average-Cost MDP: Direct Method

$$V(s) = \mathbf{T}(V) = \min_a \left\{ c(s,a) - g + \sigma_{\mathcal{P}_s^a}(V) \right\}$$

How to solve the robust Bellman equation?

- Discounted setting: apply the robust Bellman operator recursively (γ-contraction)

Robust Average-Cost MDP: Direct Method

$$V(s) = \mathbf{T}(V) = \min_a \left\{c(s,a) - g + \sigma_{\mathcal{P}_s^a}(V)\right\}$$

How to solve the robust Bellman equation?

- Discounted setting: apply the robust Bellman operator recursively (γ-contraction)
- **Average setting: not a contraction, may have multiple fixed points and algorithm may diverge**

Robust Average-Cost MDP: Direct Method

$$V(s) = \mathbf{T}(V) = \min_a \left\{ c(s,a) - g + \sigma_{\mathcal{P}_s^a}(V) \right\}$$

How to solve the robust Bellman equation?

- Discounted setting: apply the robust Bellman operator recursively (γ-contraction)
- **Average setting: not a contraction, may have multiple fixed points and algorithm may diverge**
- Robust relative value iteration (RRVI)
  - subtract an offset function to keep iterates stable
  - prove it is multi-step contraction

Robust Average-Cost MDP: Direct Method

Robust relative value iteration

INPUT: V_0, ϵ and arbitrary $s^* \in \mathcal{S}$
WHILE TRUE
FOR all $s \in \mathcal{S}$
$V_{t+1}(s) \leftarrow \min_a(c(s,a) + \sigma_{\mathcal{P}_s^a}(V_t) - f(V_t))$
OUTPUT: $f(V_t), V_t$

- For example $f(V) = V(s^*)$ for some reference state s^* and $f(V)$ is the mean of V, to "offset" the increase of V

Robust Average-Cost MDP: Direct Method

Convergence of robust relative value iteration

$(f(V_t), V_t)$ converges to a solution to the optimal robust Bellman equation

Robust Average-Cost MDP: Direct Method

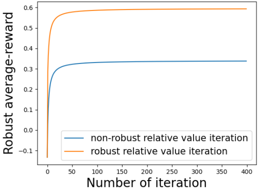
Convergence of robust relative value iteration

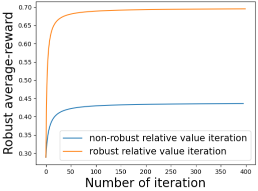
$(f(V_t), V_t)$ converges to a solution to the optimal robust Bellman equation

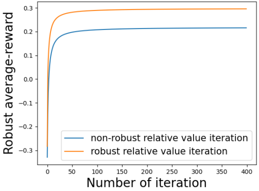
- Finds a solution to optimal robust Bellman equation and hence optimal robust average-cost and optimal policy
- Linear convergence rate, faster than limit method

Non-robust value iteration vs robust value iteration:

- under different uncertainty sets (contamination model, total variation model and KL-divergence model)
- evaluate worst-case performance of obtained policy







RRVI is more robust than non-robust relative value iteration

Model-free Method for Robust Average-Cost RL

- Idea: generalize RVI Q-learning to robust setting
- Major challenges:
  - The Bellman operator for robust average-cost MDPs is not contraction: possible multiple fixed point
  - Construction of unbiased estimator for robust Bellman operator for various uncertainty sets

Robust Q-Learning for Average RL

Optimal robust Bellman equation

If (g, Q) is a solution to the optimal robust Bellman equation

$$Q(s,a) = r(s,a) - g + \sigma_{\mathcal{P}_s^a}(V_Q), \forall s, a,$$

then 1) $g = g^*_{\mathcal{P}}$;
2) the greedy policy w.r.t. Q: $\pi_Q(s) = \arg\min_a Q(s,a)$ is an optimal robust policy;
3) $V_Q(s) \triangleq \min_a Q(s,a) = V_{\mathrm{P}}^{\pi_Q}(s) + ce$ for some $\mathrm{P} \in \Omega_g^{\pi_Q}, c \in \mathbb{R}$.

- worst-case transition kernels: $\Omega_g^{\pi} = \{\mathrm{P} \in \mathcal{P} : g_{\mathrm{P}}^{\pi} = g_{\mathcal{P}}^{\pi}\}$
- relative value function: $V_{\mathrm{P}}^{\pi} = \mathbb{E}_{\pi,\mathrm{P}}\left[\sum_{t=0}^{\infty} \mathrm{P}^t (r - g_{\mathrm{P}}^{\pi})\right]$

Robust Q-Learning for Average RL

Optimal robust Bellman equation

If (g, Q) is a solution to the optimal robust Bellman equation

$$Q(s,a) = r(s,a) - g + \sigma_{\mathcal{P}_s^a}(V_Q), \forall s, a,$$

then 1) $g = g^*_{\mathcal{P}}$;
2) the greedy policy w.r.t. Q: $\pi_Q(s) = \arg\min_a Q(s,a)$ is an optimal robust policy;
3) $V_Q(s) \triangleq \min_a Q(s,a) = V_{\mathsf{P}}^{\pi_Q}(s) + ce$ for some $\mathsf{P} \in \Omega_g^{\pi_Q}, c \in \mathbb{R}$.

- worst-case transition kernels: $\Omega_g^{\pi} = \{\mathsf{P} \in \mathcal{P} : g_{\mathsf{P}}^{\pi} = g_{\mathcal{P}}^{\pi}\}$
- relative value function: $V_{\mathsf{P}}^{\pi} = \mathbb{E}_{\pi,\mathsf{P}}\left[\sum_{t=0}^{\infty} \mathsf{P}^t (r - g_{\mathsf{P}}^{\pi})\right]$
- non-robust setting: Bellman equation is linear, and thus structure of solutions can be easily characterized
- robust Bellman equation is non-linear

Robust Q-Learning for Average RL

Robust RVI Q-learning

INPUT: Q_0, α_n, N

FOR $n = 0, ..., N-1$

$$Q_{n+1} \leftarrow Q_n + \alpha_n \big(\hat{\mathbf{H}} Q_n - f(Q_n) - Q_n\big)$$

OUTPUT: Q_N

- $\hat{\mathbf{H}}Q$: unbiased estimator of $\mathbf{H}Q = c(s,a) + \sigma_{\mathcal{P}_s^a}(V_Q)$
- $f(Q) : \mathbb{R}^{|\mathcal{S}\mathcal{A}|} \to \mathbb{R}$: "offset" increase of Q_n and keep iterates stable
- $f(Q_n)$: estimator of average-cost $g^*_{\mathcal{P}}$

Robust Q-Learning for Average RL

Convergence of Robust Q-Learning

If $\hat{\mathbf{H}}$ is unbiased and has bounded variance, then almost surely,
1) $f(Q_n)$ converges to $g^*_{\mathcal{P}}$;
2) greedy policy $\pi_{Q_n}(s) \triangleq \arg\max_a Q_n(s, a)$ converges to an optimal robust average-cost policy.

- To show convergence, we need structure of solution to robust average-cost Bellman equation to characterize the equilibrium of associated ODE, and prove it is globally asymptotically stable

Robust Q-Learning for Robust Average-Cost RL

- How to construct $\hat{\mathbf{H}}$?
  - R-contamination model: MLE method

$$\mathcal{P}_s^a = \{(1-R)p_s^a + Rq | q \in \Delta_{\mathcal{S}}\}, \text{ for some } 0 \leq R \leq 1$$

  - Other uncertainty models, e.g., total variation,Chi-square,Wasserstein distance?
- The support function $\sigma_{\mathcal{P}}(V)$ w.r.t. general uncertainty sets is non-linear in nominal kernel
- MLE method $\Rightarrow$ biased estimator

Robust Q-Learning for Average RL

- Multi-level Monte-Carlo method (Blanchet and Glynn, 2015)

For any s, a:

- Generate N according to $\textbf{Geo}(\Psi)$
- Sample 2^{N+1} samples: $\{s'_i\}, i = 1, ..., 2^{N+1}$
- divide these 2^{N+1} samples into two groups: samples with odd indices, and samples with even indices
- individually calculate the empirical distribution of s' using the even-index samples, odd-index ones, all the samples, and the first sample: $\hat{\mathsf{P}}^{a,E}_{s,N+1} = \frac{1}{2^N}\sum_{i=1}^{2^N} \mathbb{1}_{s'_{2i}}, \quad \hat{\mathsf{P}}^{a,O}_{s,N+1} = \frac{1}{2^N}\sum_{i=1}^{2^N} \mathbb{1}_{s'_{2i-1}}, \hat{\mathsf{P}}^{a}_{s,N+1} = \frac{1}{2^{N+1}}\sum_{i=1}^{2^{N+1}} \mathbb{1}_{s'_{i}}, \quad \hat{\mathsf{P}}^{a,1}_{s,N+1} = \mathbb{1}_{s'_1}$
- Use these estimated transition kernels as nominal kernels to construct four estimated uncertainty sets $\hat{\mathcal{P}}^{a,E}_{s,N+1}, \hat{\mathcal{P}}^{a,O}_{s,N+1}, \hat{\mathcal{P}}^{a}_{s,N+1}, \hat{\mathcal{P}}^{a,1}_{s,N+1}$

The multi-level estimator is then defined as

$$\hat{\sigma}_{\mathcal{P}^a_s}(V) \triangleq \sigma_{\hat{\mathcal{P}}^{a,1}_{s,N+1}}(V) + \frac{\Delta_N(V)}{p_N}, \tag{1}$$

where $p_N = \Psi(1-\Psi)^N$ and

$$\Delta_N(V) \triangleq \sigma_{\hat{\mathcal{P}}^a_{s,N+1}}(V) - \frac{\sigma_{\hat{\mathcal{P}}^{a,E}_{s,N+1}}(V) + \sigma_{\hat{\mathcal{P}}^{a,O}_{s,N+1}}(V)}{2}.$$

Robust Q-Learning for Average RL

For uncertainty sets including contamination model, total variation, Chi-squared divergence, Kullback-Leibler (KL) divergence and Wasserstein distance:

$$\mathbb{E}[\hat{\sigma}_{\mathcal{P}_s^a}(V)] = \sigma_{\mathcal{P}_s^a}(V),$$

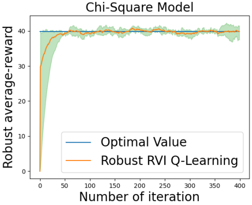
$$\mathrm{Var}[\hat{\sigma}_{\mathcal{P}_s^a}(V)(s)] \leq C(1 + \|V\|^2).$$

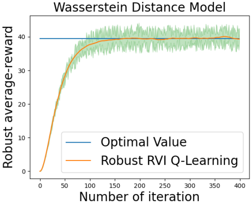
- For five uncertainty sets above, $\hat{\sigma}_{\mathcal{P}_s^a}(V)$ is unbiased and has bounded variance
- Implies convergence of robust RVI Q-learning
- Can also be applied to robust discounted setting

Experiments

Convergence of robust Q-learning

- Different uncertainty sets, e.g., Chi-Square Model and Wasserstein distance model
- Plot $f(Q_t)$ (estimate of average reward)
- Baseline is computed using model-based RVI method discussed before





Robust Q-learning converges to the optimal robust average-reward

Outline

Summary

- Robust average-cost RL
- Fundamental understanding
  - Robust average-cost Bellman equation
  - Solution characterization
  - Blackwell optimality
- Model-based approach
  - Limit method
  - Direct method
- Model-free approach: robust RVI Q-learning with convergence guarantee

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*.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2021). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76.

Asadi, K. and Littman, M. L. (2017). An alternative softmax operator for reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, volume 70, pages 243–252. JMLR.

Atkeson, C. G. and Morimoto, J. (2003). Nonparametric representation of policies and value functions: A trajectory-based approach. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 1643–1650.

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

Reference II

Bhandari, J. and Russo, D. (2021). On the linear convergence of policy gradient methods for finite MDPs. In *Proc. International Conference on Artifical Intelligence and Statistics (AISTATS)*, pages 2386–2394. PMLR.

Blanchet, J. H. and Glynn, P. W. (2015). Unbiased monte carlo for optimization and functions of expectations via multi-level randomization. In *2015 Winter Simulation Conference (WSC)*, pages 3656–3667. IEEE.

Bolte, J., Daniilidis, A., and Lewis, A. (2007). The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. (2021). Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*.

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*.

Reference III

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)*.

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

Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the Polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer.

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

Kruger, A. Y. (2003). On Fréchet subdifferentials. *Journal of Mathematical Sciences*, 116(3):3325–3358.

Reference IV

Lin, Y.-C., Hong, Z.-W., Liao, Y.-H., Shih, M.-L., Liu, M.-Y., and Sun, M. (2017). Tactics of adversarial attack on deep reinforcement learning agents. In *Proc. International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 3756–3762.

Liu, Z., Bai, Q., Blanchet, J., Dong, P., Xu, W., Zhou, Z., and Zhou, Z. (2022). Distributionally robust q-learning. In *International Conference on Machine Learning*, pages 13623–13643. PMLR.

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.

Morimoto, J. and Doya, K. (2005). Robust reinforcement learning. *Neural computation*, 17(2):335–359.

Reference V

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.

Panaganti, K., Xu, Z., Kalathil, D., and Ghavamzadeh, M. (2022). Robust reinforcement learning using offline data. *arXiv preprint arXiv:2208.05129*.

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.

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.

Puterman, M. L. (1994). Markov decision processes: Discrete stochastic dynamic programming.

Reference VI

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.

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.

Tewari, A. and Bartlett, P. L. (2007). Bounded parameter markov decision processes with average reward criterion. In *International Conference on Computational Learning Theory*, pages 263–277. Springer.

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*.

Reference VII

Wang, Y., Velasquez, A., Atia, G., Prater-Bennette, A., and Zou, S. (2023a). Model-free robust average-reward reinforcement learning. In *submitted to ICML*.

Wang, Y., Velasquez, A., Atia, G., Prater-Bennette, A., and Zou, S. (2023b). Robust average-reward Markov decision processes. In *Proc. Conference on Artificial Intelligence (AAAI)*.

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

Wang, Y. and Zou, S. (2022). Policy gradient method for robust reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, volume 162, pages 23484–23526. PMLR.

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

Reference VIII

Yang, W., Zhang, L., and Zhang, Z. (2021). Towards theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics. *arXiv preprint arXiv:2105.03863*.

Zhou, Z., Bai, Q., Zhou, Z., Qiu, L., Blanchet, J., and Glynn, P. (2021). Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *Proc. International Conference on Artifical Intelligence and Statistics (AISTATS)*, pages 3331–3339. PMLR.

Outline

Model-Free Robust Discounted RL

- Policy evaluation: robust TD (tabular), robust TDC (with function approximation)
- Optimal control: robust Q-learning (value-based), robust policy gradient (policy-based)

Value-Based Optimal Control: Robust Q-Learning

- Goal: Find a policy optimizing the worst-case performance

$$Q^{\pi}_{\mathcal{P},\gamma}(s,a) = \max_{\kappa \in \bigotimes_{t\geq 0}\mathcal{P}} \mathbb{E}_{\kappa}\left[\sum_{t=0}^{\infty}\gamma^t c(S_t,A_t)|S_0 = s, A_0 = a, \pi\right]$$

$$Q^{*}_{\mathcal{P},\gamma}(s,a) = \min_{\pi}\max_{\kappa \in \bigotimes_{t\geq 0}\mathcal{P}} \mathbb{E}_{\kappa}\left[\sum_{t=0}^{\infty}\gamma^t c(S_t,A_t)|S_0 = s, A_0 = a, \pi\right]$$

- Finding π^* is equivalent to find $Q^*_{\mathcal{P},\gamma}$

Value-Based Optimal Control: Robust Q-Learning

- **Robust Bellman operator** (Nilim and El Ghaoui, 2004): $\mathbf{T}Q(s,a) = c(s,a) + \gamma\sigma_{\mathcal{P}_s^a}(\min_{a\in\mathcal{A}} Q(s,a))$, where $\sigma_{\mathcal{P}}(v) = \max_{p\in\mathcal{P}} p^\top v$
- $\mathbf{T}$ is a γ-contraction and has $Q^*_{\mathcal{P},\gamma}$ as its unique fixed point: $\mathbf{T}Q^* = Q^*$
- Idea: recursively apply $\mathbf{T}$

Value-Based Optimal Control: Robust Q-Learning

- **Robust Bellman operator** (Nilim and El Ghaoui, 2004):
  $\mathbf{T}Q(s,a) = c(s,a) + \gamma\sigma_{\mathcal{P}_s^a}(\min_{a\in\mathcal{A}} Q(s,a))$
- Idea: recursively apply $\mathbf{T}$
- Model-free setting:
  - No information about the environment or the uncertainty set $\mathcal{P}$
  - Samples are generated under the nominal environment, generally is different from the worst-case environment
- Estimated the support function $\sigma_{\mathcal{P}_s^a}(Q)$ using the nominal samples

R-Contamination Uncertainty Sets

In this work, we mainly focus on R-contamination uncertainty set:

- $\mathcal{P}_s^a = \left\{(1-R)p_s^a + Rq | q \in \Delta_{|\mathcal{S}|}\right\}, s \in \mathcal{S}, a \in \mathcal{A},$ for some $0 \leq R \leq 1$
- Adversarial model: nature can arbitrarily modify transition kernel with probability R

Design of Robust Q-Learning

For nominal sample $O_t = (s_t, a_t, s_{t+1})$:

- Maximum likelihood estimation of transition kernel $\hat{p}_t \triangleq \mathbb{1}_{s_{t+1}}$
- Estimated uncertainty set $\hat{\mathcal{P}}_t \triangleq \left\{(1-R)\hat{p}_t + Rq | q \in \Delta_{|\mathcal{S}|}\right\}$
- Compute the support function w.r.t. $\hat{\mathcal{P}}_t$:
  $\sigma_{\hat{\mathcal{P}}_t}(V_t) = (1-R)V_t(s_{t+1}) + R\max_s V_t(s)$
- Update Q-function
  $Q_{t+1}(s_t, a_t) \leftarrow (1-\alpha_t)Q_t(s_t, a_t) + \alpha_t(c_t + \gamma\sigma_{\hat{\mathcal{P}}_t}(\min_a Q_t))$

Robust Q-learning

Initialization: T, $Q_0(s,a)$ for all (s,a), behavior policy π_b, s_0, step size α_t
For $t = 0, 1, 2, ..., T-1$
  Choose a_t according to $\pi_b(\cdot|s_t)$
  Observe s_{t+1} and c_t
  $V_t(s) \leftarrow \min_{a \in \mathcal{A}} Q_t(s,a),\ \forall s \in \mathcal{S}$
  $Q_{t+1}(s_t,a_t) \leftarrow (1-\alpha_t)Q_t(s_t,a_t) + \alpha_t(c_t + \gamma\sigma_{\hat{\mathcal{P}}_t}(V_t))$
  $Q_{t+1}(s,a) \leftarrow Q_t(s,a)$ for $(s,a) \neq (s_t,a_t)$
Output: Q_T

Theoretical Results

Theorem

(Asymptotic Convergence) If step sizes α_t *satisfy that* $\sum_{t=0}^{\infty} \alpha_t = \infty$ *and* $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$*, then* $Q_t \to Q^*_{\mathcal{P},\gamma}$ *as* $t \to \infty$ *almost surely.*

Theoretical Results

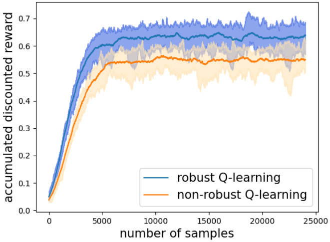
Assumption: The Markov chain induced by behavior policy π_b and transition kernel p_s^a is uniformly ergodic

Theorem

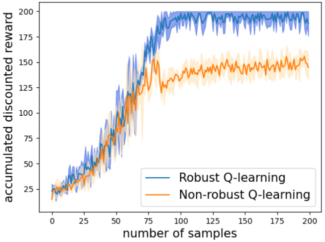
(Finite-Time Error Bound) For any ϵ, set $T = \tilde{\mathcal{O}}\left(\frac{1}{\mu_{\min}(1-\gamma)^5\epsilon^2} + \frac{t_{mix}}{\mu_{\min}(1-\gamma)}\right)$, then $\|Q_T - Q^*_{\mathcal{P},\gamma}\| \leq \mathcal{O}(\epsilon)$.

- Matches the sample complexity of non-robust Q-learning (up to some constants)
- First online, model-free method for robust RL with sample complexity result

Train Q-learning and robust Q-learning under a uniformly perturbed MDP
Test their outputs in the real unperturbed environment
Robust Q-learning achieves higher reward than Q-learning



(a) FrozenLake



(b) Cartpole

Summary on Robust Q-Learning

- For R-contamination model, use maximum likelihood estimation as the estimated nominal transition kernel, and define the estimated uncertainty set
- The support function w.r.t. the estimated uncertainty set is unbiased
- This method can be also applied to policy evaluation problem, e.g., robust TD (tabular case) or robust TDC (function approximation case)

Policy-Based Optimal Control: Robust Policy Gradient

Value-based method:

- Obtains the optimal policy using the robust value functions as an intermediate step, not direct
- Has great memory cost when the problem scale is large

Our work: 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 $V^{\pi}_{\mathcal{P},\gamma}$ may not be differentiable and non-convex
$V^{\pi}_{\mathcal{P},\gamma}(s) = \max_{\kappa \in \otimes_{t\geq 0}\mathcal{P}} \mathbb{E}_{\kappa}\left[\sum_{t=0}^{\infty} \gamma^t c(S_t, A_t) | S_0 = s, \pi\right]$

- Generalize the vanilla policy gradient to the robust policy sub-gradient method, which shows 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

Main Contributions

- Derivation of robust policy gradient: $\partial V^{\pi_\theta}_{\mathcal{P},\gamma}(s)$
- Global optimality guarantee and finite-time complexity bound
- Model-free robust actor-critic, its convergence and sample complexity

Robust Policy Gradient

- Idea: derive gradient of $J_\rho(\pi) \triangleq \mathbb{E}_\rho[V^\pi_{\mathcal{P},\gamma}(S)]$, and run gradient descent

Robust Policy Gradient

- Idea: derive gradient of $J_\rho(\pi) \triangleq \mathbb{E}_\rho[V^\pi_{\mathcal{P},\gamma}(S)]$, and run gradient descent
- Robust value function $V^\pi_{\mathcal{P},\gamma}$ is not differentiable everywhere because of max over κ

$$V^\pi_{\mathcal{P},\gamma}(s) = \max_\kappa \mathbb{E}_\kappa \left[\sum_{t=0}^{\infty} \gamma^t c(S_t, A_t) | S_0 = s, \pi \right]$$

Robust Policy Gradient

- Idea: derive gradient of $J_\rho(\pi) \triangleq \mathbb{E}_\rho[V^\pi_{\mathcal{P},\gamma}(S)]$, and run gradient descent
- Robust value function $V^\pi_{\mathcal{P},\gamma}$ is not differentiable everywhere because of max over κ

$$V^\pi_{\mathcal{P},\gamma}(s) = \max_\kappa \mathbb{E}_\kappa \left[\sum_{t=0}^{\infty} \gamma^t c(S_t, A_t) | S_0 = s, \pi \right]$$

- Major challenge lies in the max operator

Robust Policy Sub-gradient

- Consider a parametric policy class $\Pi_\Theta = \{\pi_\theta : \theta \in \Theta\}$

Theorem (Robust Policy Sub-gradient)

Define

$$\psi_\rho(\theta) \triangleq \frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)} \sum_{s\in\mathcal{S}} d_{s_\theta}^{\pi_\theta}(s) \sum_{a\in\mathcal{A}} \nabla\pi_\theta(a|s) Q_{\mathcal{P},\gamma}^{\pi_\theta}(s,a)$$
$$+ \frac{1}{1-\gamma+\gamma R} \sum_{s\in\mathcal{S}} d_{\rho}^{\pi_\theta}(s) \sum_{a\in\mathcal{A}} \nabla\pi_\theta(a|s) Q_{\mathcal{P},\gamma}^{\pi_\theta}(s,a),$$

then (1) almost everywhere in Θ, $J_\rho(\theta)$ is differentiable and $\psi_\rho(\theta) = \nabla J_\rho(\theta)$;
(2) at non-differentiable θ, $\psi_\rho(\theta) \in \partial J_\rho(\theta)$.

- $\partial J_\rho(\theta)$: set of Fréchet sub-differential (Kruger, 2003) of J_ρ at θ

Robust Policy Sub-gradient

- Consider a parametric policy class $\Pi_\Theta = \{\pi_\theta : \theta \in \Theta\}$

Theorem (Robust Policy Sub-gradient)

Define

$$\psi_\rho(\theta) \triangleq \frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)} \sum_{s\in\mathcal{S}} d_{s_\theta}^{\pi_\theta}(s) \sum_{a\in\mathcal{A}} \nabla\pi_\theta(a|s) Q_{\mathcal{P},\gamma}^{\pi_\theta}(s,a)$$
$$+ \frac{1}{1-\gamma+\gamma R} \sum_{s\in\mathcal{S}} d_\rho^{\pi_\theta}(s) \sum_{a\in\mathcal{A}} \nabla\pi_\theta(a|s) Q_{\mathcal{P},\gamma}^{\pi_\theta}(s,a),$$

then (1) almost everywhere in Θ, $J_\rho(\theta)$ is differentiable and $\psi_\rho(\theta) = \nabla J_\rho(\theta)$;
(2) at non-differentiable θ, $\psi_\rho(\theta) \in \partial J_\rho(\theta)$.

- $\partial J_\rho(\theta)$: set of Fréchet sub-differential (Kruger, 2003) of J_ρ at θ
- Reduces to vanilla policy gradient if $R = 0$

Robust Policy Sub-gradient Algorithm

Input: T, α_t
Initialization: θ_0
FOR $t = 0, 1, ..., T-1$
$\theta_{t+1} \leftarrow \prod_{\Theta}(\theta_t - \alpha_t \psi_\mu(\theta_t))$
Output: θ

Robust Policy Sub-gradient Algorithm

Input: T, α_t
Initialization: θ_0
FOR $t = 0, 1, ..., T-1$
$\theta_{t+1} \leftarrow \prod_{\Theta}(\theta_t - \alpha_t \psi_\mu(\theta_t))$
Output: θ

- Vanilla policy gradient is able to find globally optimal policy for non-robust RL, e.g., (Bhandari and Russo, 2021; Agarwal et al., 2021; Cen et al., 2021)

Robust Policy Sub-gradient Algorithm

Input: T, α_t
Initialization: θ_0
FOR $t = 0, 1, ..., T-1$
$\theta_{t+1} \leftarrow \prod_{\Theta}(\theta_t - \alpha_t \psi_\mu(\theta_t))$
Output: θ

- Vanilla policy gradient is able to find globally optimal policy for non-robust RL, e.g., (Bhandari and Russo, 2021; Agarwal et al., 2021; Cen et al., 2021)
- Question: is robust policy sub-gradient able to converge to global optimum of $J_\rho(\theta)$?

Robust Policy Sub-gradient Algorithm

Input: T, α_t
Initialization: θ_0
FOR $t = 0, 1, ..., T - 1$
$\theta_{t+1} \leftarrow \prod_{\Theta}(\theta_t - \alpha_t \psi_\mu(\theta_t))$
Output: θ

- Vanilla policy gradient is able to find globally optimal policy for non-robust RL, e.g., (Bhandari and Russo, 2021; Agarwal et al., 2021; Cen et al., 2021)
- Question: is robust policy sub-gradient able to converge to global optimum of $J_\rho(\theta)$?
- Answer: Yes!

Convex-Like: PL-Condition

PL-condition (Karimi et al., 2016; Bolte et al., 2007):

Theorem (PL-Condition)

Under direct policy parameterization,

$$J_\rho(\theta) - J_\rho^* \leq C_{PL} \max_{\hat{\pi} \in (\Delta(\mathcal{A}))^{|\mathcal{S}|}} \langle \pi_\theta - \hat{\pi}, \psi_\rho(\theta) \rangle .$$

Robust Policy Sub-gradient: Global Optimality

Theorem (Global Optimality under Direct Parameterization)

If $\alpha_t > 0$, $\sum_{t=0}^{\infty} \alpha_t = \infty$ *and* $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$, *then under direct policy parameterization,* θ_T *converges to a global optimum of* $J_\rho(\theta)$ *as* $T \to \infty$ *almost surely.*

Robust Policy Sub-gradient: Global Optimality

Theorem (Global Optimality under Direct Parameterization)

If $\alpha_t > 0$, $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$, then under direct policy parameterization, θ_T converges to a global optimum of $J_\rho(\theta)$ as $T \to \infty$ almost surely.

- Sub-gradient method converges to stationary points: $\{\theta : 0 \in \partial J_\rho(\theta)\}$
- Stationary point is globally optimal due to PL-condition

Outline

4 Robust Discounted RL

5 Robust Sub-gradient

Smoothed Robust Policy Gradient

Robust policy sub-gradient method:

- Complexity is generally difficult to establish

Smoothed Robust Policy Gradient

Robust policy sub-gradient method:

- Complexity is generally difficult to establish

Solution: smoothed robust policy gradient

Smoothed Robust Policy Gradient

Smoothed robust Bellman operator:

$$\mathbf{T}_{\sigma}^{\pi} V(s) = \mathbb{E}_{A \sim \pi(\cdot|s)} \left[c(s, A) + \gamma(1 - R) \sum_{s' \in \mathcal{S}} p_{s,s'}^{A} V(s') + \gamma R \cdot \mathrm{LSE}(\sigma, V) \right],$$

where $\mathrm{LSE}(\sigma, V) = \frac{\log(\sum_{i=1}^{d} e^{\sigma V(i)})}{\sigma}$ for $V \in \mathbb{R}^d$ and some $\sigma > 0$

Smoothed Robust Policy Gradient

Smoothed robust Bellman operator:

$$\mathbf{T}_\sigma^\pi V(s) = \mathbb{E}_{A\sim\pi(\cdot|s)}\left[c(s,A) + \gamma(1-R)\sum_{s'\in\mathcal{S}} p_{s,s'}^A V(s') + \gamma R \cdot \mathrm{LSE}(\sigma, V)\right],$$

where $\mathrm{LSE}(\sigma, V) = \frac{\log(\sum_{i=1}^d e^{\sigma V(i)})}{\sigma}$ for $V \in \mathbb{R}^d$ and some $\sigma > 0$

- $\mathrm{LSE}(\sigma, V)$ converges to $\max_s V(s)$ as $\sigma \to \infty$

Smoothed Robust Policy Gradient

Smoothed robust Bellman operator:

$$\mathbf{T}_\sigma^\pi V(s) = \mathbb{E}_{A\sim\pi(\cdot|s)}\left[c(s,A) + \gamma(1-R)\sum_{s'\in\mathcal{S}} p_{s,s'}^A V(s') + \gamma R \cdot \mathrm{LSE}(\sigma, V)\right],$$

where $\mathrm{LSE}(\sigma, V) = \frac{\log(\sum_{i=1}^d e^{\sigma V(i)})}{\sigma}$ for $V \in \mathbb{R}^d$ and some $\sigma > 0$

- $\mathrm{LSE}(\sigma, V)$ converges to $\max_s V(s)$ as $\sigma \to \infty$
- T_σ^π is a contraction, V_σ^π is the fixed point of T_σ^π
  softmax will not induce contraction (Asadi and Littman, 2017)

Smoothed Robust Policy Gradient

Smoothed robust Bellman operator:

$$\mathbf{T}_\sigma^\pi V(s) = \mathbb{E}_{A\sim\pi(\cdot|s)}\left[c(s,A) + \gamma(1-R)\sum_{s'\in\mathcal{S}} p_{s,s'}^A V(s') + \gamma R \cdot \mathrm{LSE}(\sigma, V)\right],$$

where $\mathrm{LSE}(\sigma, V) = \frac{\log(\sum_{i=1}^d e^{\sigma V(i)})}{\sigma}$ for $V \in \mathbb{R}^d$ and some $\sigma > 0$

- $\mathrm{LSE}(\sigma, V)$ converges to $\max_s V(s)$ as $\sigma \to \infty$
- T_σ^π is a contraction, V_σ^π is the fixed point of T_σ^π
  softmax will not induce contraction (Asadi and Littman, 2017)
- V_σ^π is differentiable in θ and converges to V^π as $\sigma \to \infty$

Smoothed Robust Policy Gradient

- $J_\rho^\sigma(\theta) = \sum_{s \in \mathcal{S}} \rho(s) V_\sigma^{\pi_\theta}(s)$: smoothed robust objective

Smoothed Robust Policy Gradient

- $J_\rho^\sigma(\theta) = \sum_{s \in \mathcal{S}} \rho(s) V_\sigma^{\pi_\theta}(s)$: smoothed robust objective
- Gradient of $J_\rho^\sigma(\theta)$:

$$\nabla J_\rho^\sigma(\theta) = B(\rho, \theta) + \frac{\gamma R \sum_{s \in \mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)} B(s, \theta)}{(1-\gamma) \sum_{s \in \mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}},$$

where $B(s, \theta) \triangleq \frac{1}{1-\gamma+\gamma R} \sum_{s' \in \mathcal{S}} d_s^\pi(s') \sum_{a \in \mathcal{A}} \nabla \pi_\theta(a|s') Q_\sigma^{\pi_\theta}(s', a)$, and $B(\rho, \theta) \triangleq \mathbb{E}_{S \sim \rho}[B(S, \theta)]$.

- Smoothed robust policy gradient: $\theta_{t+1} \leftarrow \prod_\Theta(\theta_t - \alpha_t \nabla J_\rho^\sigma(\theta))$

Smoothed Robust Policy Gradient

- $J_\rho^\sigma(\theta) = \sum_{s\in\mathcal{S}} \rho(s) V_\sigma^{\pi_\theta}(s)$: smoothed robust objective
- Gradient of $J_\rho^\sigma(\theta)$:

$$\nabla J_\rho^\sigma(\theta) = B(\rho,\theta) + \frac{\gamma R \sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)} B(s,\theta)}{(1-\gamma)\sum_{s\in\mathcal{S}} e^{\sigma V_\sigma^{\pi_\theta}(s)}},$$

where $B(s,\theta) \triangleq \frac{1}{1-\gamma+\gamma R} \sum_{s'\in\mathcal{S}} d_s^\pi(s') \sum_{a\in\mathcal{A}} \nabla \pi_\theta(a|s') Q_\sigma^{\pi_\theta}(s',a)$, and $B(\rho,\theta) \triangleq \mathbb{E}_{S\sim\rho}[B(S,\theta)]$.

- Smoothed robust policy gradient: $\theta_{t+1} \leftarrow \prod_\Theta(\theta_t - \alpha_t \nabla J_\rho^\sigma(\theta))$

Even though gradient is for J_ρ^σ, the algorithm can still find a global optimum of J_ρ by choosing a large σ

Global optimality and Complexity

Consider direct policy parameterization

Theorem

For any $\epsilon > 0$*, set* $\sigma = \mathcal{O}(\epsilon^{-1})$ *and* $T = \mathcal{O}(\epsilon^{-3})$*, then*

$$\min_{t \leq T-1} J(\theta_t) - J^* \leq 3\epsilon.$$

Global optimality and Complexity

Consider direct policy parameterization

Theorem

For any $\epsilon > 0$*, set* $\sigma = \mathcal{O}(\epsilon^{-1})$ *and* $T = \mathcal{O}(\epsilon^{-3})$*, then*

$$\min_{t \leq T-1} J(\theta_t) - J^* \leq 3\epsilon.$$

- If $R = 0$, i.e., no robustness is considered, complexity reduces to $\mathcal{O}(\epsilon^{-2})$, which matches with vanilla policy gradient in (Agarwal et al., 2021)

Model-free Robust Actor-Critic

- Recall robust policy subgradient:

$$\psi_\rho(\theta) \triangleq \frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)} \sum_{s \in \mathcal{S}} d_{s_\theta}^{\pi_\theta}(s) \sum_{a \in \mathcal{A}} \nabla \pi_\theta(a|s) Q_{\mathcal{P},\gamma}^{\pi_\theta}(s,a) + \frac{1}{1-\gamma+\gamma R} \sum_{s \in \mathcal{S}} d_{\rho}^{\pi_\theta}(s) \sum_{a \in \mathcal{A}} \nabla \pi_\theta(a|s) Q_{\mathcal{P},\gamma}^{\pi_\theta}(s,a)$$

- $Q^{\pi_\theta}(s,a)$ measures cost under worst-case transition kernel and π_θ, however, only samples from simulator are available

Model-free Robust Actor-Critic

- Recall robust policy subgradient:

$$\psi_\rho(\theta) \triangleq \frac{\gamma R}{(1-\gamma)(1-\gamma+\gamma R)} \sum_{s\in\mathcal{S}} d_{s_\theta}^{\pi_\theta}(s) \sum_{a\in\mathcal{A}} \nabla\pi_\theta(a|s) Q_{\mathcal{P},\gamma}^{\pi_\theta}(s,a) + \frac{1}{1-\gamma+\gamma R} \sum_{s\in\mathcal{S}} d_\rho^{\pi_\theta}(s) \sum_{a\in\mathcal{A}} \nabla\pi_\theta(a|s) Q_{\mathcal{P},\gamma}^{\pi_\theta}(s,a)$$

- $Q^{\pi_\theta}(s,a)$ measures cost under worst-case transition kernel and π_θ, however, only samples from simulator are available

Monte Carlo does not work

Critic: Robust TD

- Parametric robust action value function Q_ζ, e.g., linear function approximation or neural network.

Input: T_c, π, β_t
Initialization: ζ, s_0
  Choose $a_0 \sim \pi(\cdot|s_0)$
FOR $t = 0, 1, ..., T_c - 1$
  Observe c_t, s_{t+1}
  Choose $a_{t+1} \sim \pi(\cdot|s_{t+1})$
  $V_t^* \leftarrow \max_s \left\{\sum_{a\in\mathcal{A}} \pi(a|s) Q_\zeta(s,a)\right\}$
  $\delta_t \leftarrow Q_\zeta(s_t, a_t) - (\underbrace{c_t + \gamma(1-R)Q_\zeta(s_{t+1}, a_{t+1}) + \gamma R V_t^*}_{\text{robust target}})$ (robust TD error)
  $\zeta \leftarrow \zeta - \beta_t \delta_t \nabla_\zeta Q_\zeta(s_t, a_t)$
Output: ζ

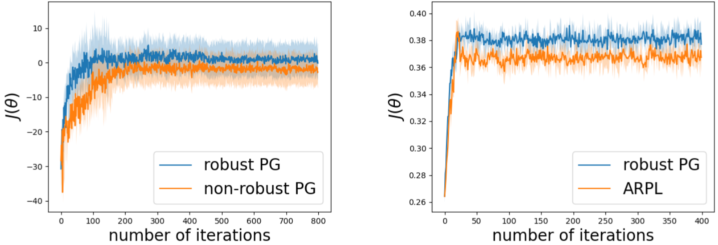
Robust Actor-Critic Algorithm

- Using robust TD algorithm to estimate robust Q-function in (smoothed) robust policy gradient
- Under tabular setting, global optimality can be established, overall sample complexity is $\mathcal{O}(\epsilon^{-7})$

Robust actor-critic algorithm can be applied with arbitrary value function/policy approximation.

Experiments

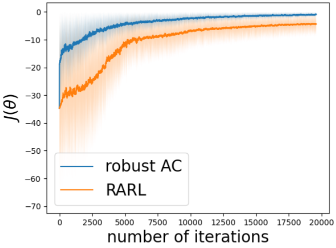
- Robust policy gradient v.s. vanilla policy gradient and ARPL Mandlekar et al. (2017)
- ARPL: Adversary randomly perturb observation then run vanilla policy gradient method using these perturbed samples
- Training on an unperturbed MDP, and evaluation on the worst-case transition kernel in $\mathcal{P}$

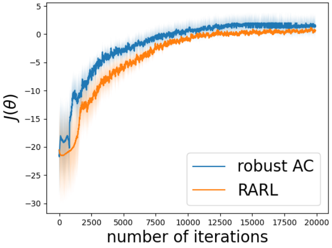


- Our robust policy gradient achieves higher reward on the worst-case transition kernel

Experiments

- Robust actor-critic v.s. RARL (Pinto et al., 2017)
- RARL: Adversary perturbs state transition. Agent and adversary are updated alternatively using gradient descent ascent.
- Training on an unperturbed MDP, and evaluation on the worst-case transition kernel in $\mathcal{P}$





- Our robust actor critic achieves higher reward on the worst-case transition kernel

Summary

- Robust policy gradient with provable global optimality
- Model-free robust actor-critic algorithm
- Can be easily scaled to large/continuous problems