Natural Actor-Critic for Robust Reinforcement Learning with Function Approximation

Ruida Zhou*, Tao Liu*,
Min Cheng, Dileep Kalathil,
P. R. Kumar, Chao Tian

Paper

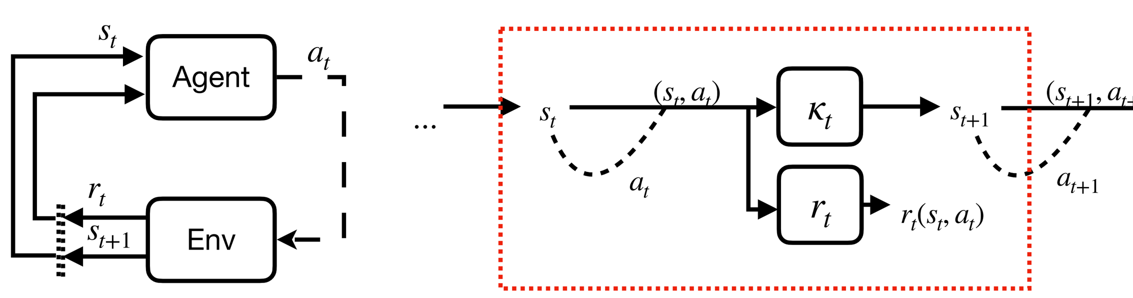
Video

NEURAL INFORMATION PROCESSING SYSTEMS

This work:

- Policy-based approach: Capable of continuous control – large action space
- Function approximation: Handling large state space – essence of deep RL
- Robust RL: Dealing with sim-to-real gap – naturally arise in application

From RL to Robust RL



MDP diagram

Unrolling MDP or RMDP

MDP $(\mathcal{S}, \mathcal{A}, \kappa, r)$ — Robust MDP $\{(\mathcal{S}, \mathcal{A}, \kappa, r) : \kappa \in \mathcal{P}^\infty\}$

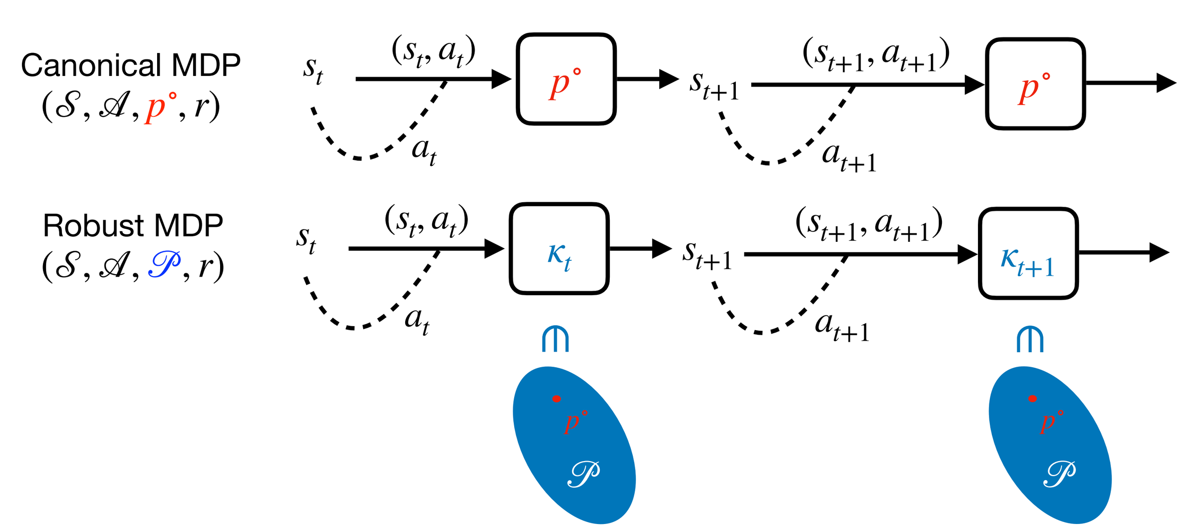
Transition $\kappa = (\kappa_0, \kappa_1, \ldots)$ typically stationary with $\kappa_t = p^\circ$ (nominal model / simulator)

Policy $a_t \sim \pi(\cdot \mid s_t)$

Value function $V^\pi_\kappa(s) = \mathbb{E}_{\kappa,\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s\right]$

Robust value $V^\pi_{\mathcal{P}}(s) = \inf_{\kappa \in \mathcal{P}^\infty} V^\pi_\kappa(s)$

Goal: Find policy to maximize $V^\pi_\kappa(\rho) = \mathbb{E}_{s\sim\rho}[V^\pi_{\mathcal{P}}(s)]$



Challenges

Key of RMDP modeling: (s, a)-rectangular uncertainty set $\mathcal{P} = \otimes_{(s,a)} \mathcal{P}_{s,a}$ facilitates dynamic programming (DP). DP is about robust Bellman operator $(\mathcal{T}^\pi_{\mathcal{P}} V)(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[r(s,a) + \gamma \inf_{p\in\mathcal{P}_{s,a}} p^\top V\right]$.

Tractable robust Bellman operator estimation $\rightsquigarrow$ efficient learning

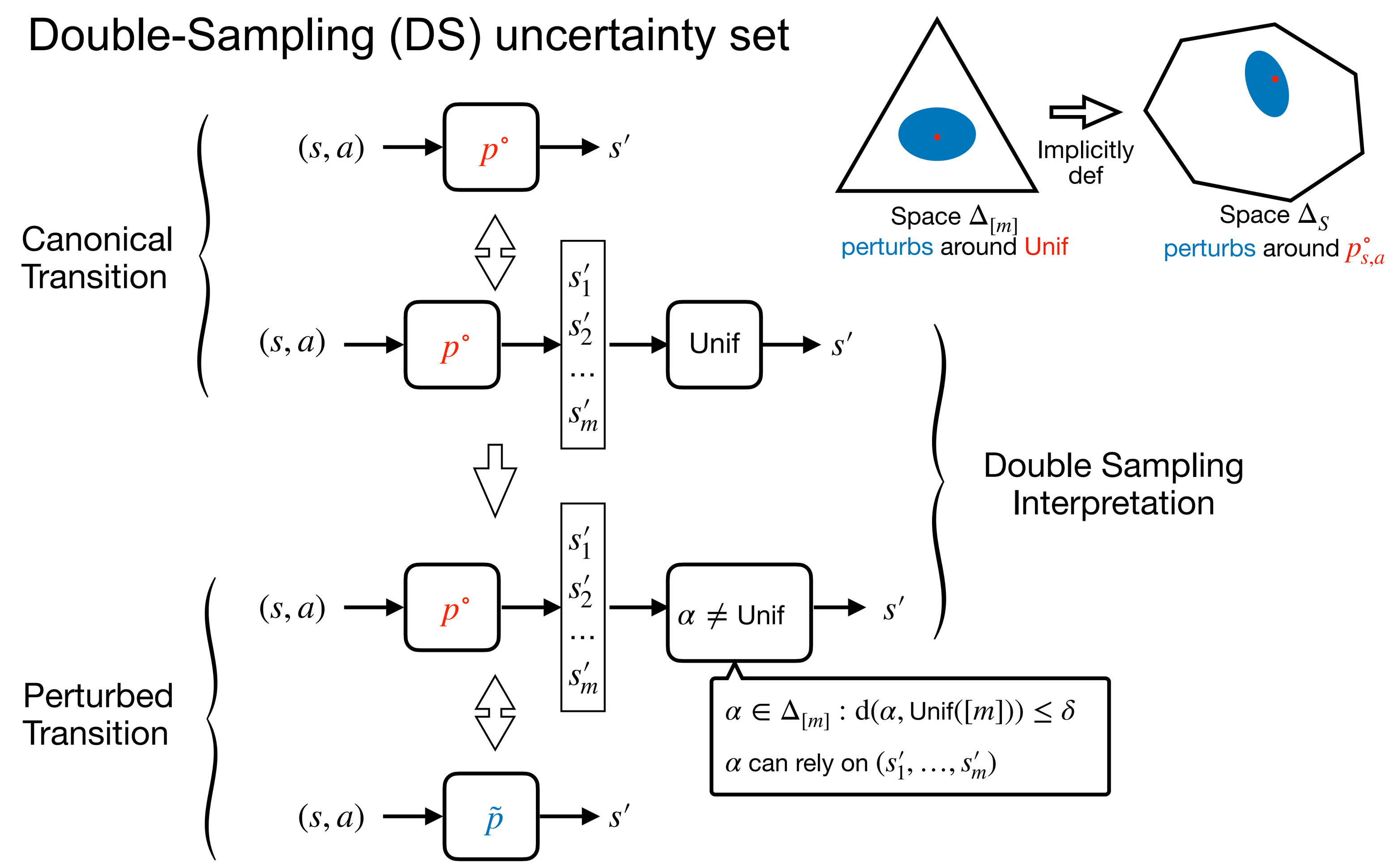
Uncertainty sets examples – facilitate robust Bellman operator estimation

- R-contamination: $\mathcal{P}_{s,a} = \{Rq + (1-R)p^\circ_{s,a} : q \in \Delta_{\mathcal{S}}\}$,
  $\inf_{p\in\mathcal{P}_{s,a}} p^\top V = (1-R)(p^\circ_{s,a})^\top V + R \min_{s'} V(s')$
- ℓ_p-norm: $\mathcal{P}_{s,a} = \{q \in \Delta_{\mathcal{S}} : \|q - p^\circ_{s,a}\|_p \le \delta\}$,
  $\inf_{p\in\mathcal{P}_{s,a}} p^\top V = (p^\circ_{s,a})^\top V - \min_{w\in\mathbb{R}} \|V - w\mathbf{1}\|_q$
- f-div: $\mathcal{P}_{s,a} = \{p \in \Delta_{\mathcal{S}} : d_f(p, p^\circ_{s,a}) = \sum_{s'} p^\circ_{s,a}(s') f(\frac{p(s')}{p^\circ_{s,a}(s')}) \le \delta\}$,
  $\inf_{p\in\mathcal{P}_{s,a}} p^\top V = \sup_{\lambda>0,\eta\in\mathbb{R}} \mathbb{E}_{s'}\left[-\lambda f^*\left(\frac{-V(s')-\eta}{\lambda}\right) - \lambda\delta - \eta\right]$

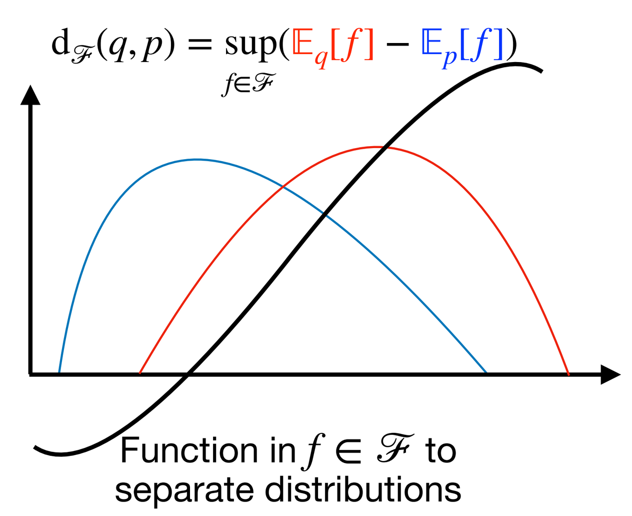
These uncertainty sets do not scale up.

Uncertainty set design

Double-Sampling (DS) uncertainty set



Integral probability metric (IPM) uncertainty set



Func class in IPM — Func approximation in RL ?!

Linear function approximation in RL:

Feature maps $\psi(s) \in \mathbb{R}^d$

Approximate value function $V_w(s) = \psi(s)^\top w$

Function class for IPM:

$\mathcal{F} = \{s \mapsto \psi(s)^\top \xi : \xi \in \mathbb{R}^d, \|\xi\| \le 1\}$

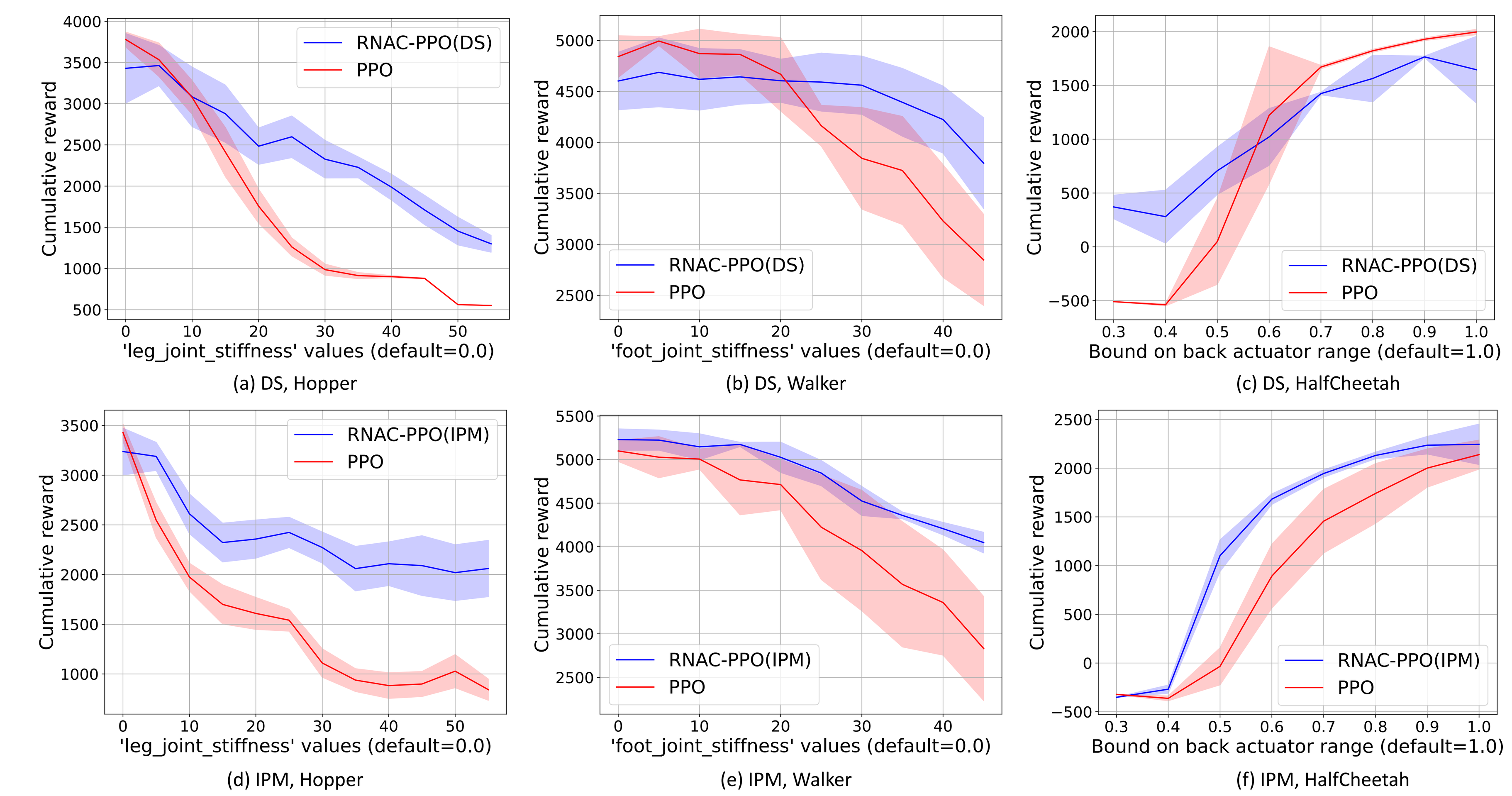
$\mathcal{P}_{s,a} = \{q : d_{\mathcal{F}}(q, p^\circ_{s,a}) = \sup_{f\in\mathcal{F}}\{q^T f - {p^\circ_{s,a}}^T f\} \le \delta\}$

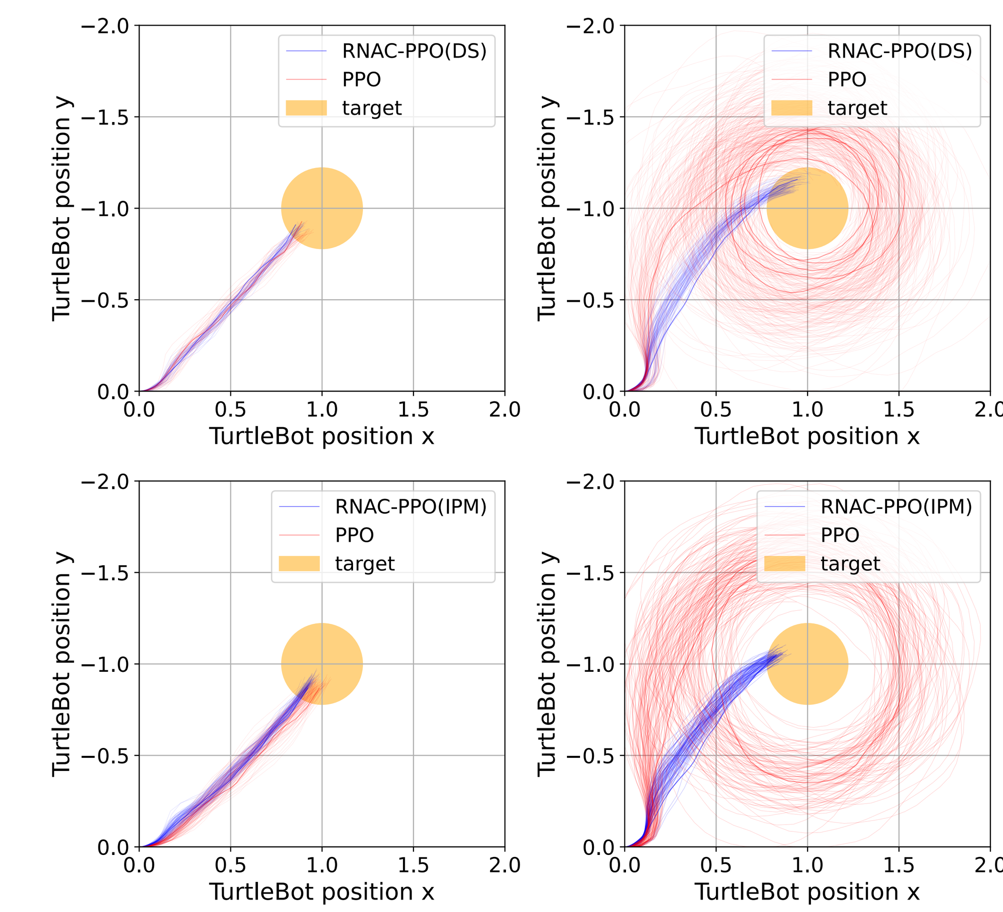
Computational tractable empirical robust Bellman operator

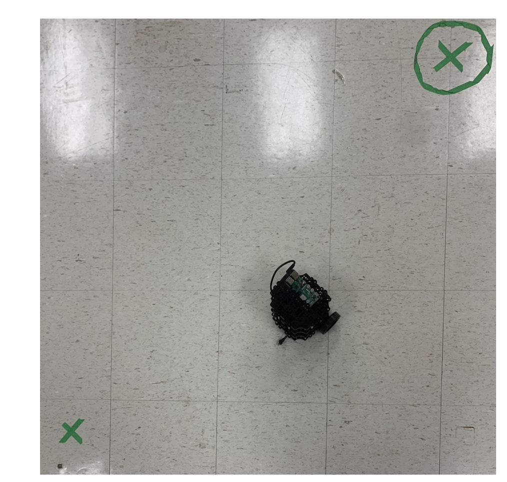
DS — $(\hat{\mathcal{T}}^\pi_{\mathcal{P}} V)(s, a, s'_{1:m}) := r(s,a) + \gamma \inf_{\alpha\in\Delta_{[m]}:d(\alpha,\text{Unif}([m]))\le\delta} \sum_{i=1}^m \alpha_i V(s'_i)$

IPM — $(\hat{\mathcal{T}}^\pi_{\mathcal{P}} V)(s, a, s'_{1:m}) := r(s,a) + \gamma V_w(s') - \gamma\delta\|w_{2:d}\|$

Experiments







We adopt a PPO actor update in the RNAC framework — RNAC-PPO;

The proposed RNAC-PPO is more robust in many simulated and real environments, including MuJoCo tasks and real TurtleBot navigation task.

Robust Natural Actor-Critic Algorithm

Algorithm 1: Robust Natural Actor-Critic

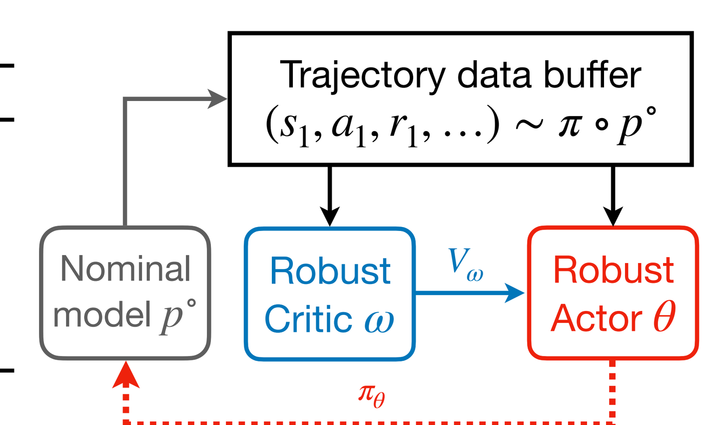
Input: $T, \eta^{0:T-1}, K, N$

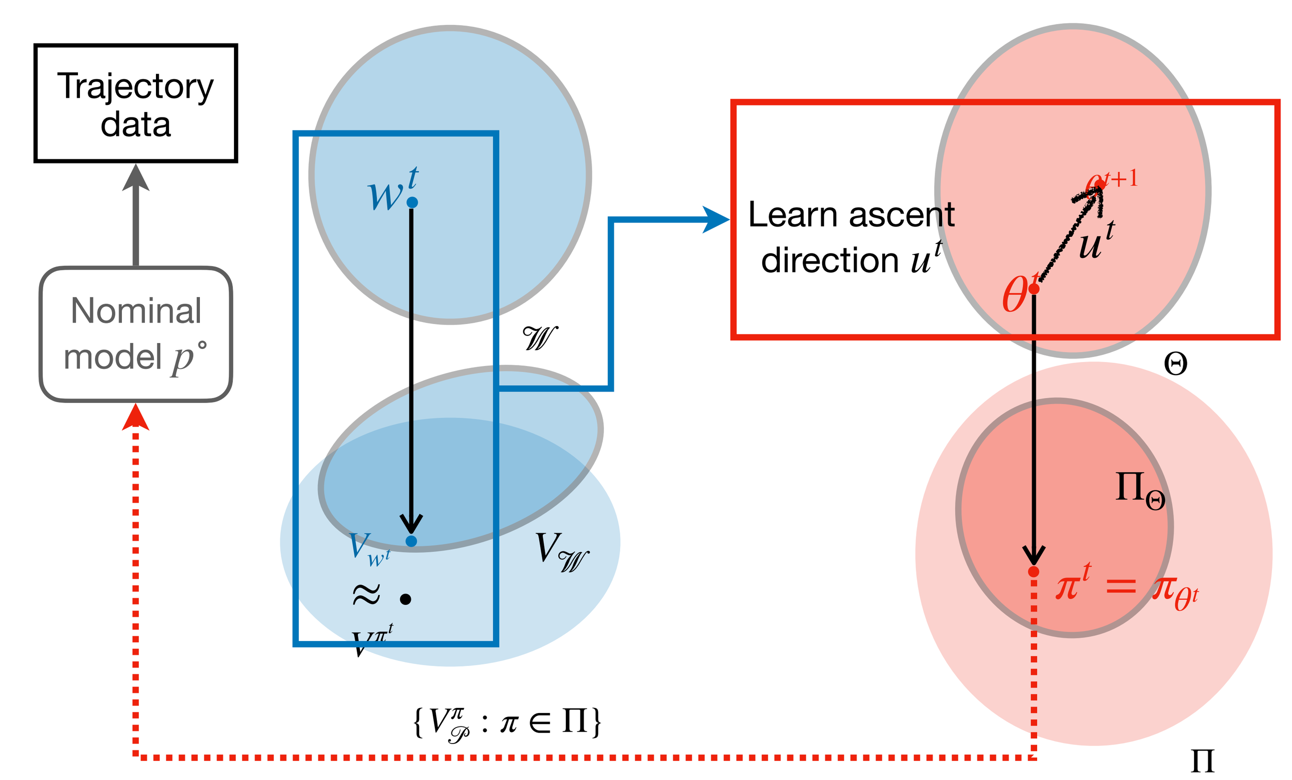
Initialize: θ^0 for policy parameterization and w_{init} for value function approximation

for $t = 0, 1, \ldots, T-1$ **do**

Robust critic updates w^t; //E.g., w^t = RLTD(π_{θ^t}, K) Algorithm 2

Robust natural actor updates θ^{t+1};//E.g., θ^{t+1} = RQNPG(θ^t, η^t, w^t, N) Algorithm 3





Theoretical Guarantee

Under linear function approximation and some assumptions, RNAC with appropriate geometrically increasing step sizes η^t, achieves $\mathbb{E}[V^{\pi^*}(\rho) - V^{\pi^T}(\rho)] = O(e^{-T}) + O(\epsilon_{stat}) + O(\epsilon_{bias})$ and an $\tilde{\mathcal{O}}(1/\epsilon^2)$ sample complexity; Same condition and constant step size, $\mathbb{E}[V^{\pi^*}(\rho) - \frac{1}{T}\sum_{t=1}^T V^{\pi^t}(\rho)] = O(1/T) + O(\epsilon_{stat}) + O(\epsilon_{bias})$ and $\tilde{\mathcal{O}}(1/\epsilon^3)$ sample complexity;

(i) $\epsilon_{stat} = \tilde{O}(1/\sqrt{N} + 1/\sqrt{K})$ is a statistical error; (ii) ϵ_{bias} is the approximation error due to limited representation power

Under general function approximation, RNAC has an $O(1/\sqrt{T})$ optimization rate and an $\tilde{\mathcal{O}}(1/\epsilon^4)$ sample complexity;