Robust Situational Reinforcement Learning in Face of Context Disturbances

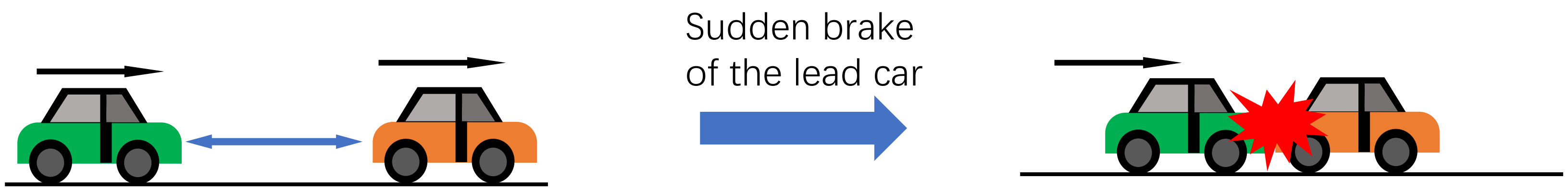
Jinpeng Zhang[1], Yufeng Zheng[2], Chuheng Zhang[3], Li Zhao[3], Lei Song[3], Yuan Zhou[1], Jiang Bian[3]

[1]Tsinghua University, [2]University of Toronto, [3]Microsoft Research Asia

Motivation

- Context variable: the dynamic and uncontrollable environmental factor in many real-world tasks
  - E.g., Inventory Control and Adaptive Cruise Control (ACC):

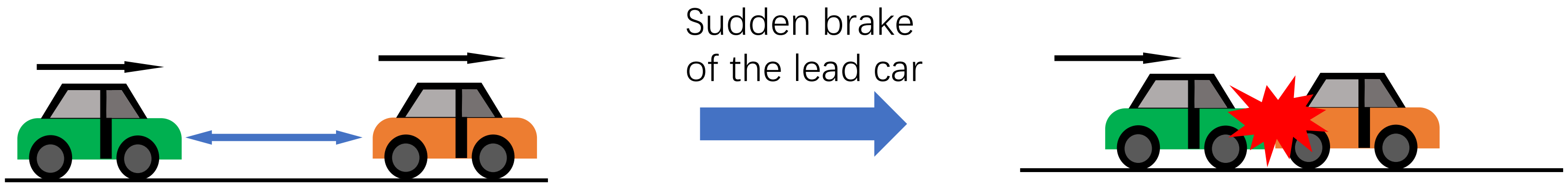
    Context variables are the customer demand and speed of lead car, respectively, which are independent of agent's action, and have large uncertainty



Motivation

- Context variable: the dynamic and uncontrollable environmental factor in many real-world tasks
  - E.g., Inventory Control and Adaptive Cruise Control (ACC):

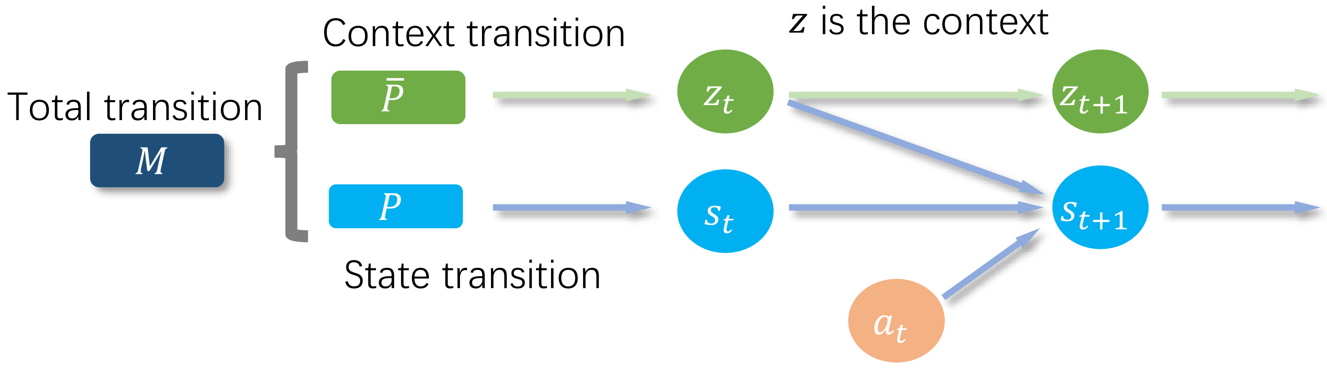
    Context variables are the customer demand and speed of lead car, respectively, which are independent of agent's action, and have large uncertainty



- Put uncertainty only to contexts!
  - After taking an action, the state of the ego car is clear
  - Robustness against worst-case context disturbances

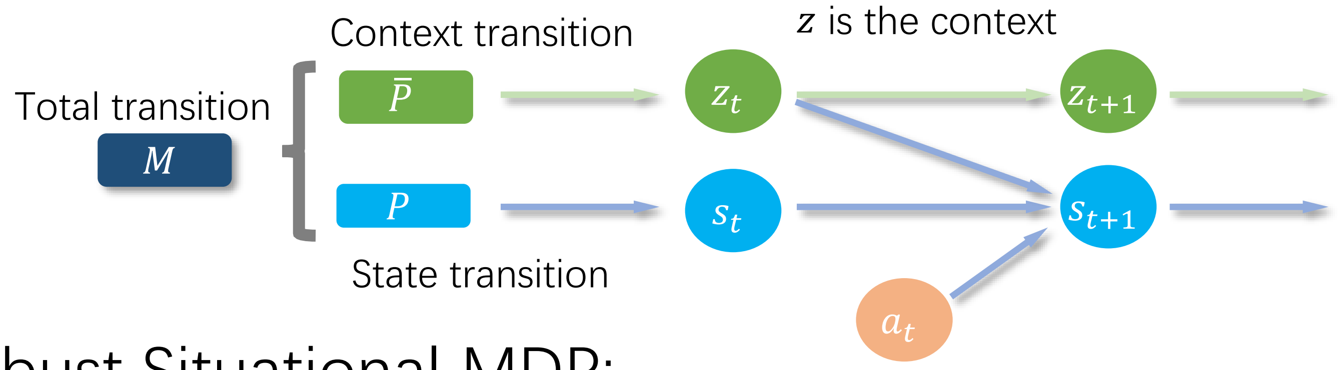
Problem Formulation

- Situational RL: factorized transitions $M(s', z'|s, z, a) = \bar{P}(z'|z)P(s'|s, z, a)$

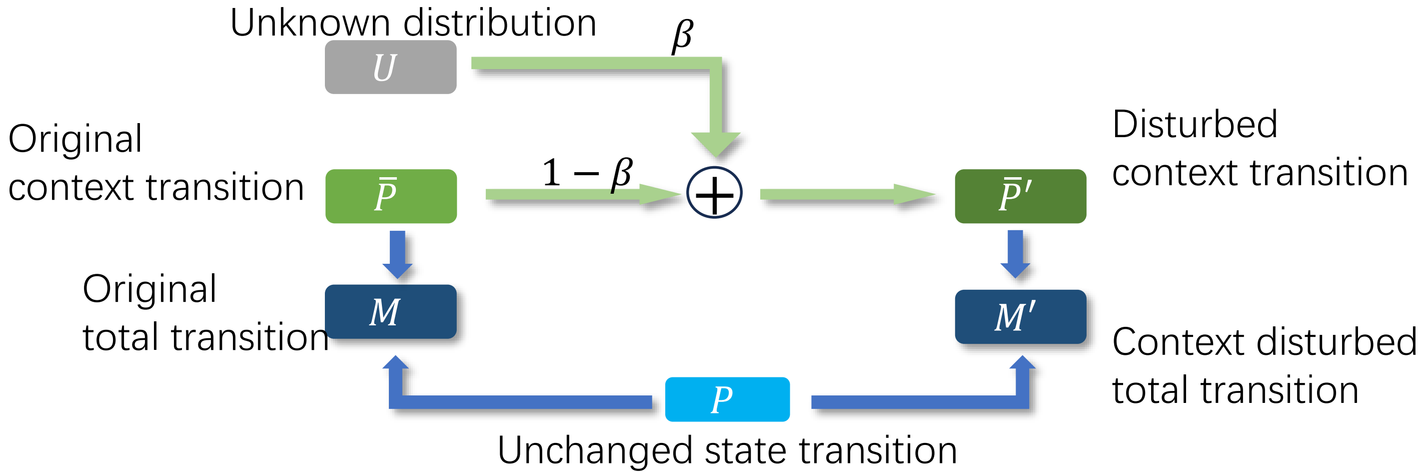


Problem Formulation

- Situational RL: factorized transitions $M(s', z'|s, z, a) = \bar{P}(z'|z)P(s'|s, z, a)$



- Robust Situational MDP:
  - Put Huber's contamination model to context transitions



Method: Basics

- Our Robust Bellman Equation

$$\mathcal{B}^{\pi}_{\text{rob}}Q(s,z,a) = r(s,z,a) + \gamma(1-\beta)\mathbb{E}_{s',z',a'}[Q(s',z',a')] + \gamma\beta \min_{U\in\Delta(\mathcal{Z})} \int_{\mathcal{Z}} \mathbb{E}_{s',a'}[Q(s',z'',a')]U(z'')dz''$$

Method: Basics

- Our Robust Bellman Equation

$$\mathcal{B}^{\pi}_{\text{rob}}Q(s,z,a) = r(s,z,a) + \gamma(1-\beta)\mathbb{E}_{s',z',a'}[Q(s',z',a')]$$
$$+\gamma\beta \underline{\min_{U\in\Delta(\mathcal{Z})}} \int_{\mathcal{Z}} \mathbb{E}_{s',a'}[Q(s',z'',a')]U(z'')dz''$$
$$= r(s,z,a) + \gamma(1-\beta)\mathbb{E}_{s',z',a'}[Q(s',z',a')]$$
$$+\gamma\beta \underline{\min_{z''}} \mathbb{E}_{s',a'}[Q(s',z'',a')]$$

Method: Basics

- Our Robust Bellman Equation

$$\mathcal{B}^{\pi}_{\text{rob}} Q(s,z,a) = r(s,z,a) + \gamma(1-\beta)\mathbb{E}_{s',z',a'}[Q(s',z',a')] + \gamma\beta \underline{\min_{U\in\Delta(\mathcal{Z})}} \int_{\mathcal{Z}} \mathbb{E}_{s',a'}[Q(s',z'',a')]U(z'')dz''$$

$$= r(s,z,a) + \gamma(1-\beta)\mathbb{E}_{s',z',a'}[Q(s',z',a')] + \gamma\beta \underline{\min_{z''}} \mathbb{E}_{s',a'}[Q(s',z'',a')]$$

- Our proposed robust Bellman equation precisely captures the setting where only deviations of the context transitions matter

Method: Deep RL Case

- To scale to large context space, we introduce the softmin smoothed robust Bellman operator

$$\mathcal{B}_\tau^\pi Q(s,z,a) = r(s,z,a) + \gamma(1-\beta)\mathbb{E}_{s',z',a'}[Q(s',z',a')] + \gamma\beta \cdot \underline{\mathrm{SoftMin}}_{z'}\left(\mathbb{E}_{s',a'}[Q(s',z',a')]\right)$$

Intuitively, as temperature $\tau \to 0$, $\mathrm{SoftMin} \to \mathrm{Min}$

Method: Deep RL Case

- To scale to large context space, we introduce the softmin smoothed robust Bellman operator

$$\mathcal{B}_\tau^\pi Q(s,z,a) = r(s,z,a) + \gamma(1-\beta)\mathbb{E}_{s',z',a'}[Q(s',z',a')] + \gamma\beta \cdot \underline{\mathrm{SoftMin}}_{z'}\left(\mathbb{E}_{s',a'}[Q(s',z',a')]\right)$$

Intuitively, as temperature $\tau \to 0$, $\mathrm{SoftMin} \to \mathrm{Min}$

- Robust Situational Soft Actor-Critic (RS-SAC):
  - The target of critic network in original SAC is changed to be the softmin smoothed robust Bellman backup

Method: Deep RL Case

- To scale to large context space, we introduce the softmin smoothed robust Bellman operator

$$\mathcal{B}^{\pi}_{\tau} Q(s,z,a) = r(s,z,a) + \gamma(1-\beta)\mathbb{E}_{s',z',a'}[Q(s',z',a')] + \gamma\beta \cdot \underline{\mathrm{SoftMin}}_{z'}\left(\mathbb{E}_{s',a'}[Q(s',z',a')]\right)$$

Intuitively, as temperature $\tau \to 0$, $\mathrm{SoftMin} \to \mathrm{Min}$

- Robust Situational Soft Actor-Critic (RS-SAC):
  - The target of critic network in original SAC is changed to be the softmin smoothed robust Bellman backup
- We theoretically show that the softmin is a reasonable approximation to the true robust Bellman equation

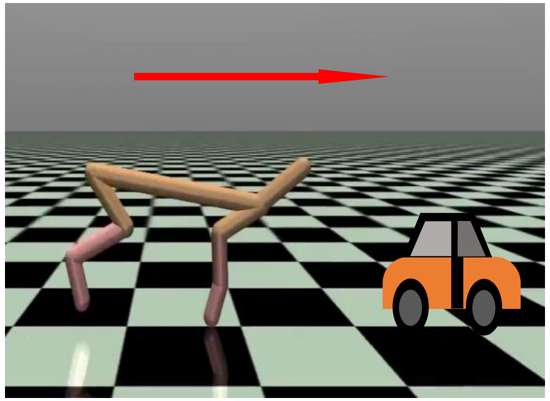
Theorem. *Let $Q_t = \mathcal{B}^{\pi}_{\tau} Q_{t-1}$ to be the t-th iteration applying the softmin smoothed robust Bellman equation and fix $\epsilon > 0$. Then there exists constant $C > 0$ such that the difference between Q_t and the true robust Q-function Q^{π}_{rob} satisfies*

$$\|Q_t - Q^{\pi}_{\mathrm{rob}}\|_{\infty} \lesssim \gamma^t \|Q_0 - Q^{\pi}_{\mathrm{rob}}\|_{\infty} + \frac{\beta C}{1-\gamma} \cdot \tau$$

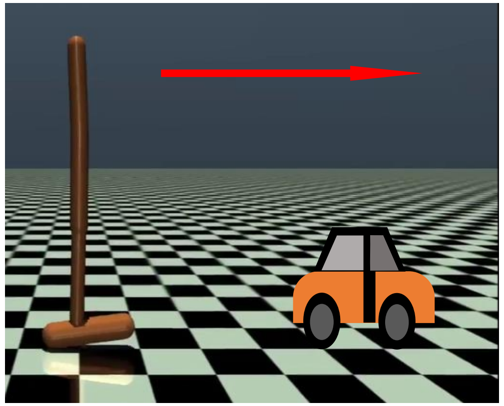
Experiments: Locomotion Control with Dynamic Contexts

- Tasks

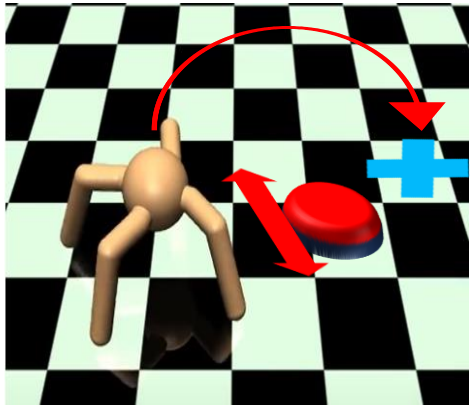
HalfCheetah-acc



Hopper-acc



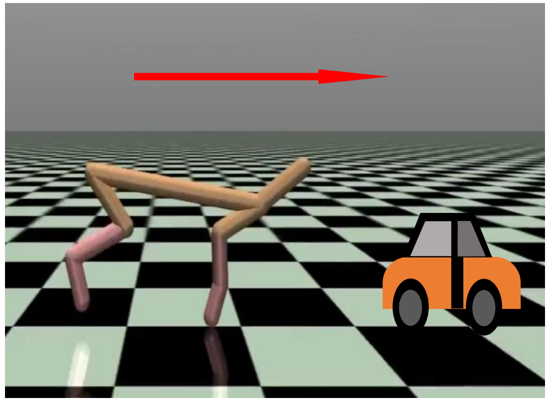
Ant-cross



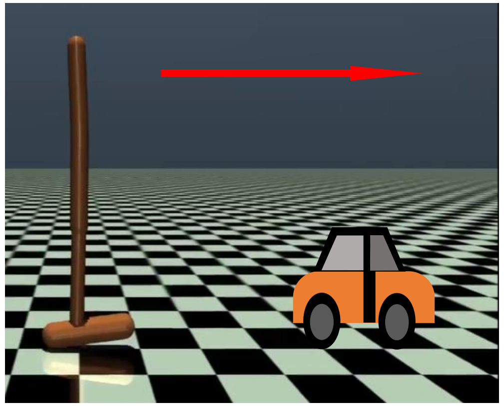
Experiments: Locomotion Control with Dynamic Contexts

- Tasks

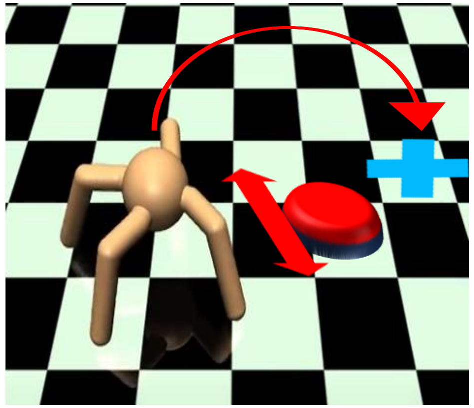
HalfCheetah-acc



Hopper-acc



Ant-cross

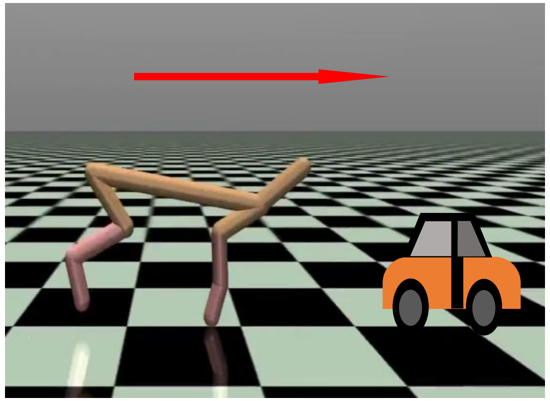


- Context transitions:
  - HalfCheetah-acc and Hopper-acc:
    - Speed of lead car $v_{t+1} = v_t + \Delta v$, where $\Delta v \sim N(\mu, \sigma)$ is the change of speed
  - Ant-cross:
    - Obstacle position $y_t \sim N(\mu, \sigma)$

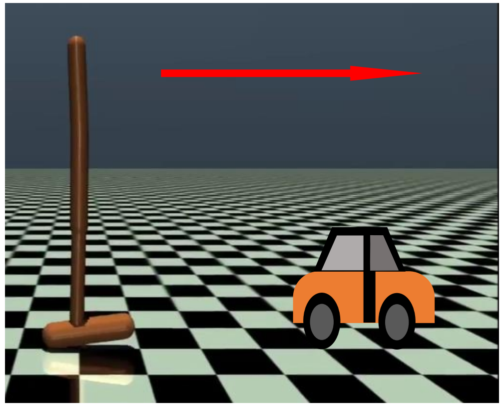
Experiments: Locomotion Control with Dynamic Contexts

- Tasks

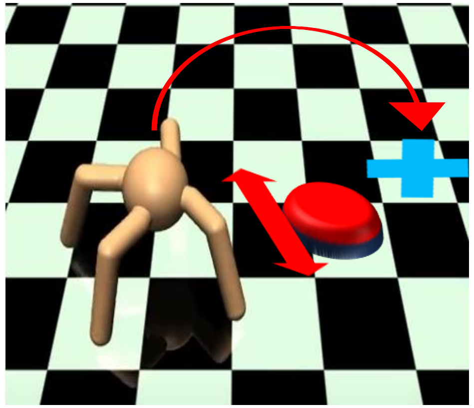
HalfCheetah-acc



Hopper-acc



Ant-cross



- Context transitions:
  - HalfCheetah-acc and Hopper-acc:
    - Speed of lead car $v_{t+1} = v_t + \Delta v$, where $\Delta v \sim N(\mu, \sigma)$ is the change of speed
  - Ant-cross:
    - Obstacle position $y_t \sim N(\mu, \sigma)$
- Will change μ and σ to test robustness against context disturbances

Experiments: Locomotion Control with Dynamic Contexts

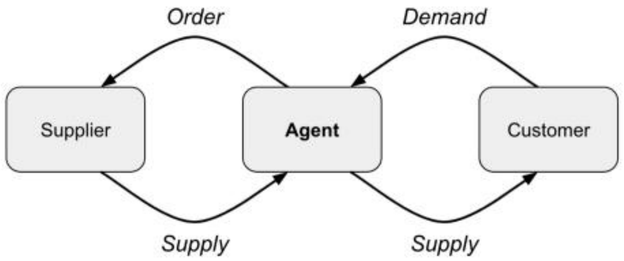
- Overall average returns

	RS-SAC	SAC	DR-SAC	PR-SAC	SC-SAC
HalfCheetah-acc	**1622.8**	1144.8	1251.6	1292.1	1474.8
Hopper-acc	**2044.8**	1921.6	1894.3	1621.8	1989.2
Ant-cross	**340.7**	**341.6**	288.5	83.7	41.3

- Our algorithm RS-SAC
    - achieves better performance in HalfCheetah-acc and Hopper-acc
    - achieves competitive performance in Ant-cross

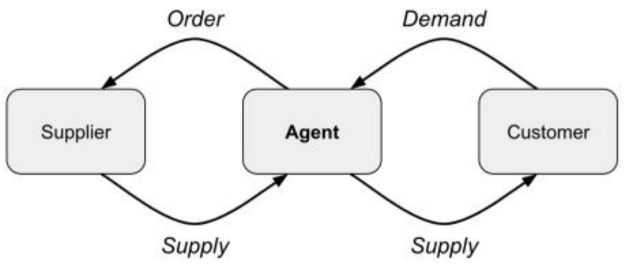
Experiments: Real-world Inventory Control

- Context variable: customer demand
  - Large uncertainty

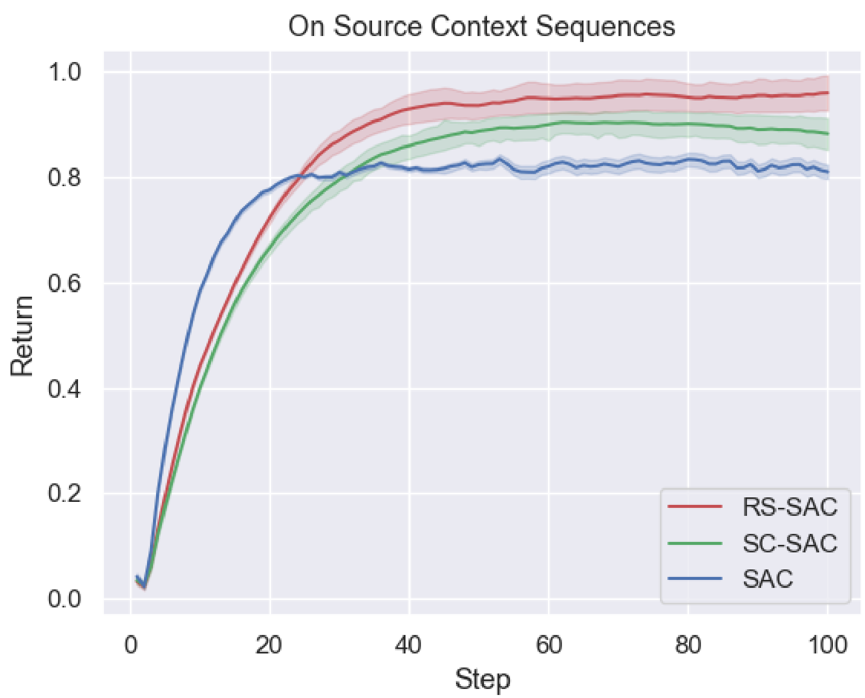


Experiments: Real-world Inventory Control

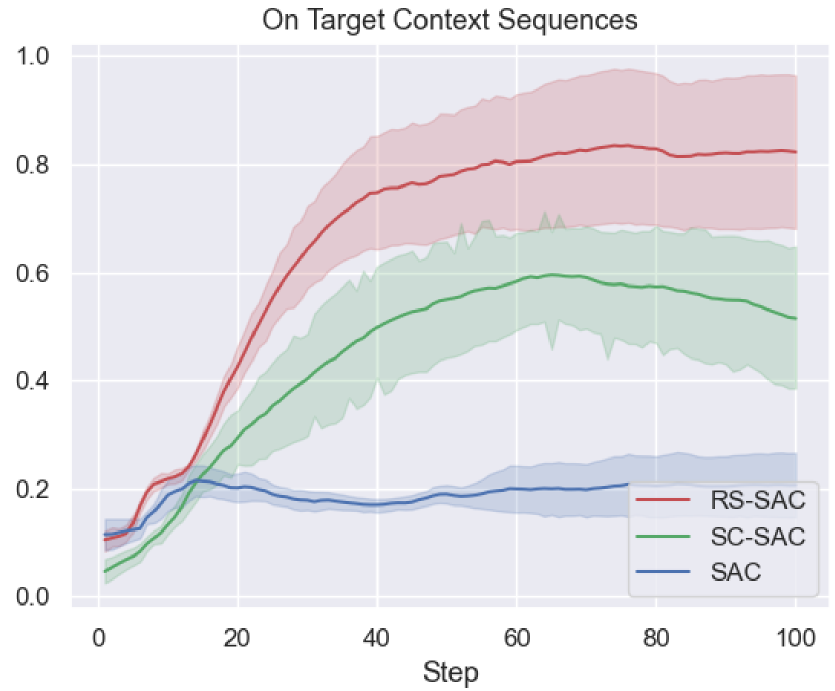
- Context variable: customer demand
  - Large uncertainty



Competitive in training



Outperform other baselines in testing



Summary

- We introduce robust situational MDP which captures the disturbances in context transitions
- We propose the softmin smoothed robust Bellman operator to apply to existing deep RL algorithms (e.g., SAC)
- Experiments on Locomotion Control tasks with dynamic contexts and inventory control tasks show that our algorithm is more robust to context disturbances