Detecting Rewards Deterioration in Episodic RL

Ido Greenberg & Shie Mannor

Technion, Israel

ICML 2021

Performance Awareness in RL

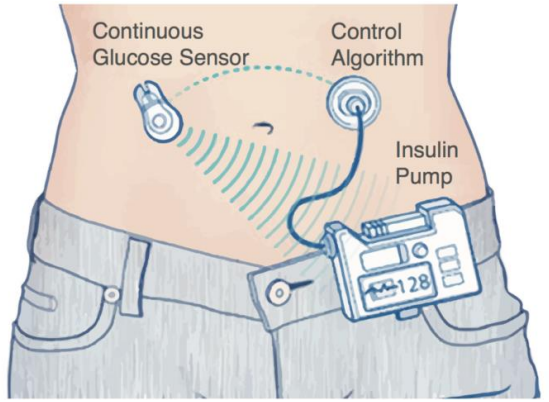
- RL in risk-intolerant applications:

train → fix → test → run in production

Most works

Our work

- Requirement: notice defects ASAP



Detecting Rewards Deterioration

- Are the rewards in recent episodes smaller than usual?

Detecting Rewards Deterioration

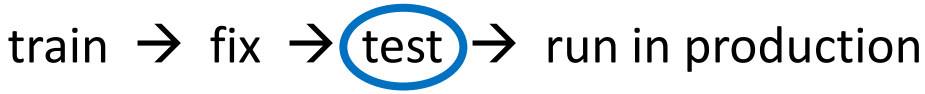
- Are the rewards in recent episodes smaller than usual?

Issue	Solution
“Smaller” than what?	

Detecting Rewards Deterioration

- Are the rewards in recent episodes smaller than usual?

Issue	Solution
“Smaller” than what?	Reference dataset (e.g. tests recordings)



Detecting Rewards Deterioration

- Are the rewards in recent episodes smaller than usual?

Issue	Solution
“Smaller” than what?	Reference dataset (e.g. tests recordings)
Rewards are not i.i.d	

- Assume i.i.d episodes
- Average reward per episode

Detecting Rewards Deterioration

- Are the rewards in recent episodes smaller than usual?

Issue	Solution
“Smaller” than what?	Reference dataset (e.g. tests recordings)
Rewards are not i.i.d	

- Assume i.i.d episodes
- Average reward per episode
  - Time-steps with small variance are more informative!

Detecting Rewards Deterioration

- Are the rewards in recent episodes smaller than usual?

Issue	Solution
“Smaller” than what?	Reference dataset (e.g. tests recordings)
Rewards are not i.i.d	

- Assume i.i.d episodes
- Average reward per episode
  - Time-steps with small variance are more informative!
- Weighted average
  - $W = \mathbf{1}^{\top}\Sigma^{-1}$
    - (all-1 vector & cov matrix)

Detecting Rewards Deterioration

- Are the rewards in recent episodes smaller than usual?

Issue	Solution
“Smaller” than what?	Reference dataset (e.g. tests recordings)
Rewards are not i.i.d	

- Assume i.i.d episodes
- Average reward per episode
  - Time-steps with small variance are more informative!
- Weighted average
  - $W = \mathbf{1}^{\top}\Sigma^{-1}$
    - (all-1 vector & cov matrix)
  - **Better** than simple average
  - Under normality: **optimal**
  - Benefit depends on Σ’s spectrum

Detecting Rewards Deterioration

- Are the rewards in recent episodes smaller than usual?

Issue	Solution
“Smaller” than what?	Reference dataset (e.g. tests recordings)
Rewards are not i.i.d	I.i.d episodes – use weighted average per episode
Sequential tests – inflation of false-alarms	

Detecting Rewards Deterioration

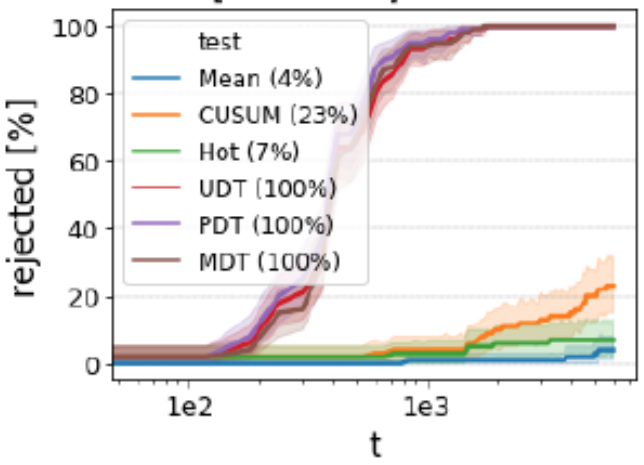
- Are the rewards in recent episodes smaller than usual?

Issue	Solution
“Smaller” than what?	Reference dataset (e.g. tests recordings)
Rewards are not i.i.d	I.i.d episodes – use weighted average per episode
Sequential tests – inflation of false-alarms	BFAR – Bootstrap for False Alarm Rate control • Sample episodes (not time-steps!) • Handle incomplete episodes

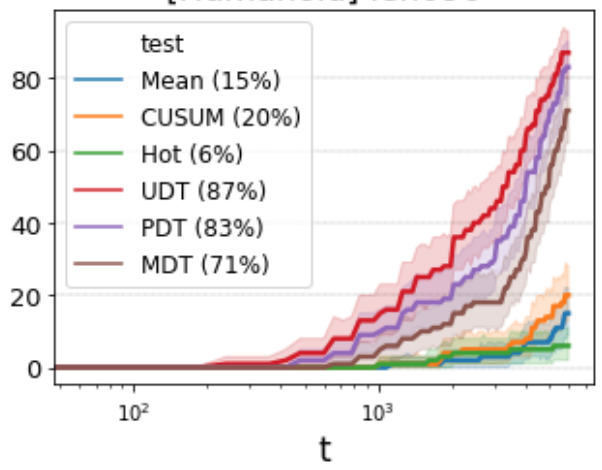
Experiments

- Probability to notice deterioration after t steps
  - Our test variants: UDT, PDT, MDT

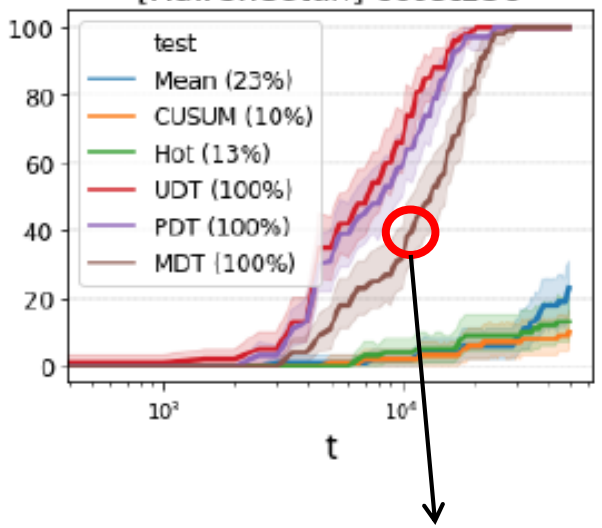
1:20 engine SNR



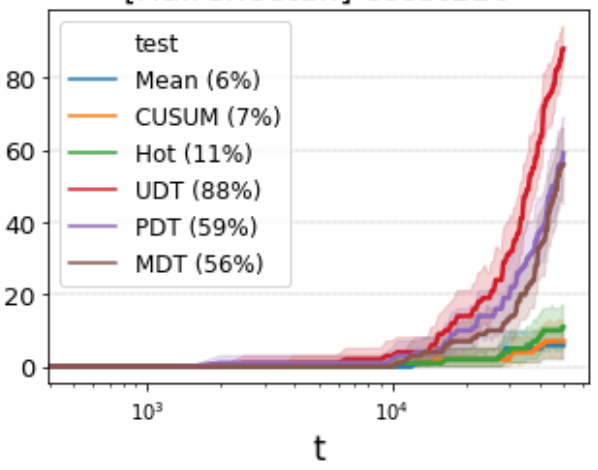
-10% leg size



+30% engine cost



+10% engine cost



"After 10^4 time-steps, MDT has 40% to detect the degradation"