Non-stationary Reinforcement Learning without Prior Knowledge: An Optimal Black-box Approach

Chen-Yu Wei Haipeng Luo
University of Southern California

Non-stationary Reinforcement Learning

Stationary RL: fixed transition / reward

expected reward of the best policy

$$\text{Regret} = \sum_{t=1}^{T} (\mu^\star - R_t)$$

learner's reward at round *t*

Non-stationary RL (our focus): time-varying transition / reward

expected reward of the best policy at round *t*

$$\text{Dynamic-Regret} = \sum_{t=1}^{T} (\mu_t^\star - R_t)$$

Non-stationary Reinforcement Learning

Measures of non-stationarity:

$$S = 1 + \sum_{t=2}^{T} \mathbf{1}\left[r_t \neq r_{t-1} \text{ or } p_t \neq p_{t-1} \right]$$

# of change points

$$V = 1 + \sum_{t=2}^{T} \left(\|r_t - r_{t-1}\|_\infty + \|p_t - p_{t-1}\|_\infty \right)$$

sum of changes between consecutive rounds

The achievable dynamic regret bound will depend on S or V.

Related Works

Popular approaches: sliding window, time discounting, periodic restarting

(Ortner et al. 2018, Mao et al., 2021, Zhou et al., 2020, Touati and Vincent, 2020, Domingues et al. 2021)

- Gives sub-optimal bound
- Requires prior knowledge on S or V (to tune hyperparameters)

Cheung, et al., (2020) developed **Bandit-over-RL**, making the above algorithms prior-knowledge free. But it worsens the bound.

Multi-armed bandit: optimal bound without knowledge of S or V

Auer, Gajane, Ortner (2019): multi-armed bandit
Chen, et al., (2019): multi-armed contextual bandit

specialized and unclear how to extend to, e.g., linear bandit, MDPs

Our Contribution

We address all these issues via a general approach that is:

- **a black-box reduction**

UCB (originally for stationary env) $\xrightarrow{\text{convert into}}$ an algorithm for non-stationary env

- **widely applicable** (bandits, episodic/infinite-horizon MDP, linear MDP, …)

- **optimal & prior-knowledge free** ~~(knowledge on S or V)~~

Applications and Comparisons (unknown S or V)

Setting	Prior Work		Our Work
Linear Bandit / Episodic tabular MDP	$S^{1/3}T^{2/3} + T^{3/4}$	$V^{1/3}T^{2/3} + T^{3/4}$	$\min\{\sqrt{ST}, V^{1/3}T^{2/3}\}$ (optimal bound)
	(Cheung et al., 2018, Fei et al., 2020, Mao et al., 2021)		
Episodic Linear MDP / Infinite-horizon MDP	$S^{1/3}T^{2/3} + T^{3/4}$	$V^{1/4}T^{3/4}$	
	(Ortner et al., 2018, Chueng et al. 2020, Zhou et al., 2020, Touati and Vincent, 2020)		
Generalized Linear Bandit	$S^{1/3}T^{2/3} + T^{3/4}$	$V^{1/5}T^{4/5}$	
	(Russac et al., 2020, Faury et al., 2021)		

- Except for (contextual) multi-armed bandits, previous work does not get $\sqrt{ST}$ even if S is known.
- For generalized linear bandits, linear MDPs, previous work does not get $V^{1/3}T^{2/3}$ even if V is known.

Algorithm

Properties of UCB (Upper-Confidence Bound) Algorithms

In the stationary environment, in every round, the algorithm can output a **scaler** $\widetilde{\mu}_t$ that is non-increasing in t, such that:

① $\widetilde{\mu}_t \geq \mu^\star$ $\mu^\star$: expected reward of the best policy

② $\sum_{\tau=1}^{t} (\widetilde{\mu}_\tau - R_\tau) \leq \widetilde{O}(\sqrt{t})$ R_τ : learner's reward in round τ

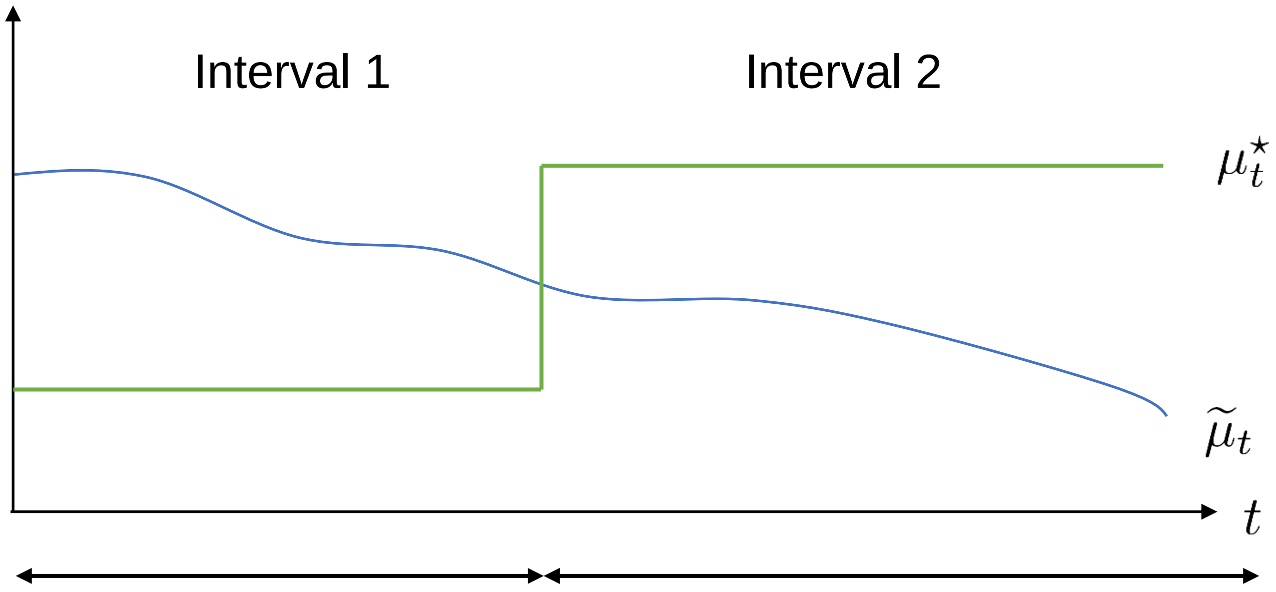
Idea of our algorithm:

Run UCB, and detect non-stationarity   ← test if ① or ② is violated

If non-stationarity is detected → restart the algorithm

<u>Testing the violation of ② is straightforward</u>

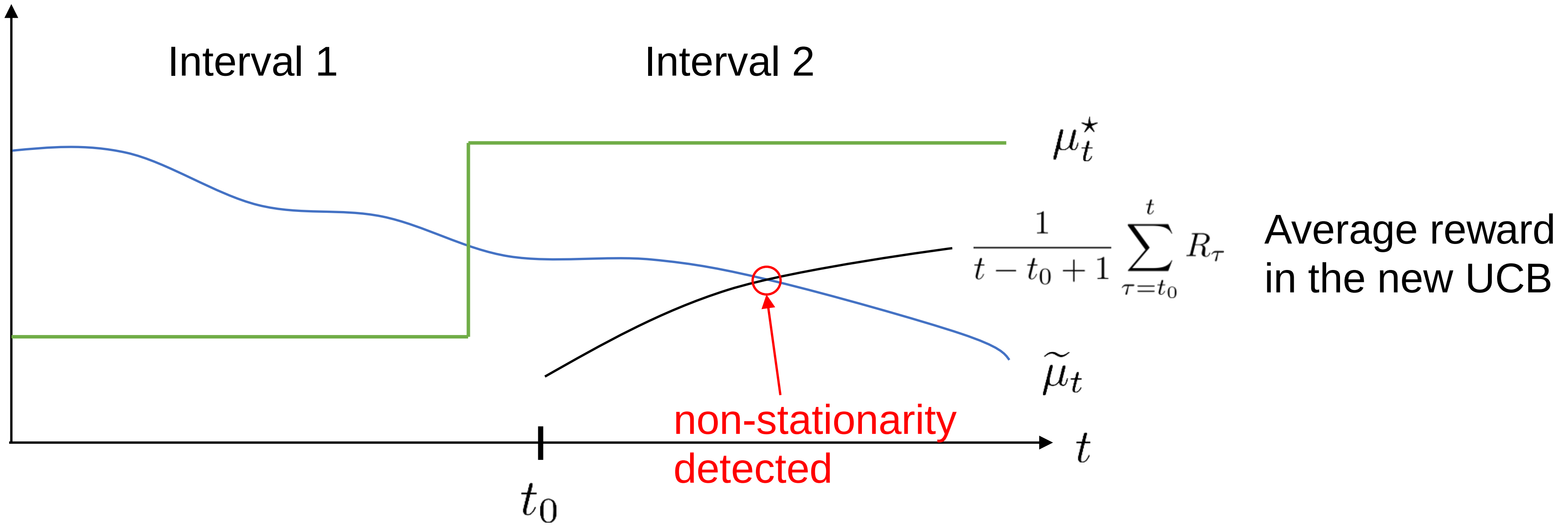
Detecting the Violation of 1 (i.e., detecting $\widetilde{\mu}_t < \mu_t^\star$)



$\widetilde{\mu}_t$ is a valid upper bound of $\mu_t^\star$

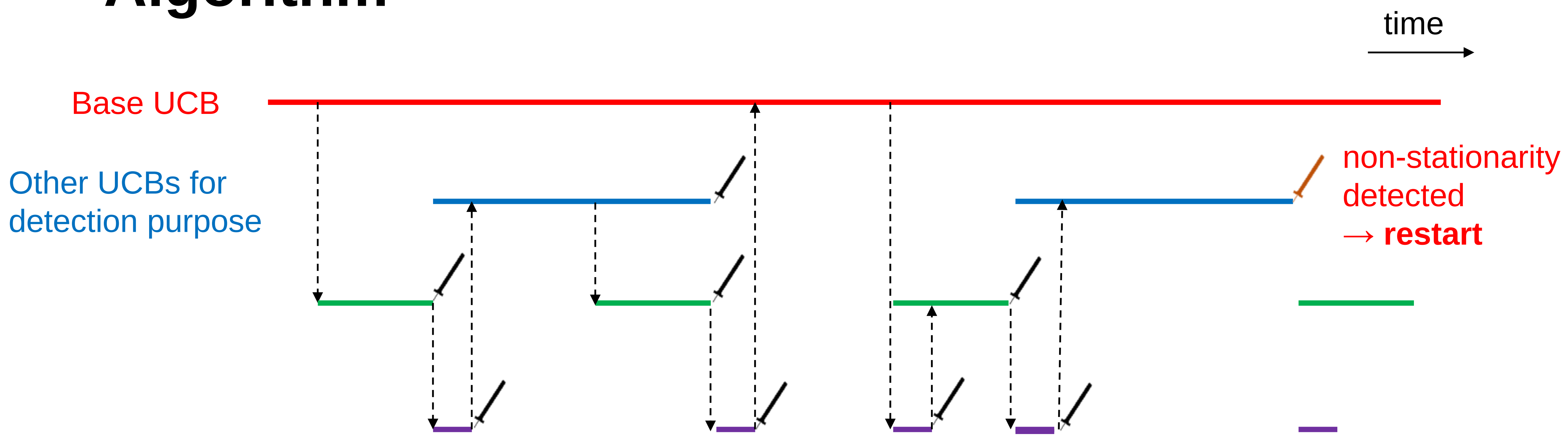
$\widetilde{\mu}_t$ is affected by staled data, and is no longer an upper bound of $\mu_t^\star$

Detecting the Violation of ① (i.e., detecting $\widetilde{\mu}_t < \mu_t^\star$)



If we start a **new** UCB algorithm from t_0 …
(assume it coexists with the original UCB)

Algorithm



- Randomly schedule **UCBs of different lengths** (shorter UCBs for quickly detecting larger change; longer UCBs for detecting smaller change)
- If multiple UCBs overlap, the shortest one has priority (the longer ones pause)
- At the end of every UCB, test if the average performance exceeds $\widetilde{\mu}_t$
  If so, restart the whole algorithm.

Summary

UCB $\sqrt{T}$

stationary RL

$\min\{\sqrt{ST}, V^{1/3}T^{2/3}\}$

non-stationary RL