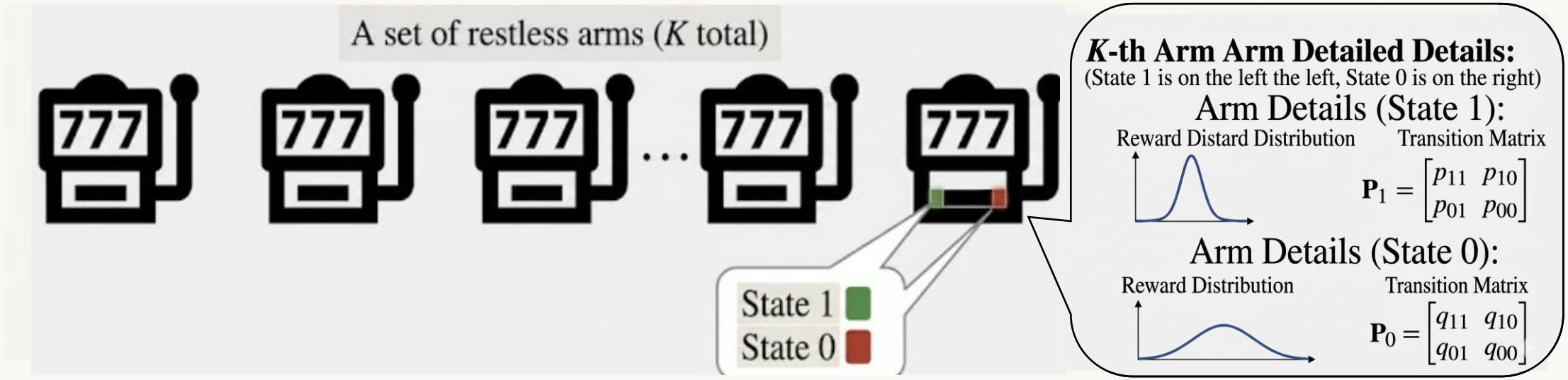
A Modularized Framework for Piecewise-Stationary Restless Bandits

Kuan-Ta Li, *Chia-Chun Lin, Ping-Chun Hsieh, and Yu-Chih Huang*

National Yang Ming Chiao Tung University

國立陽明交通大學
NATIONAL YANG MING CHIAO TUNG UNIVERSITY

Restless Multi-Armed Bandit (RMAB) Problem



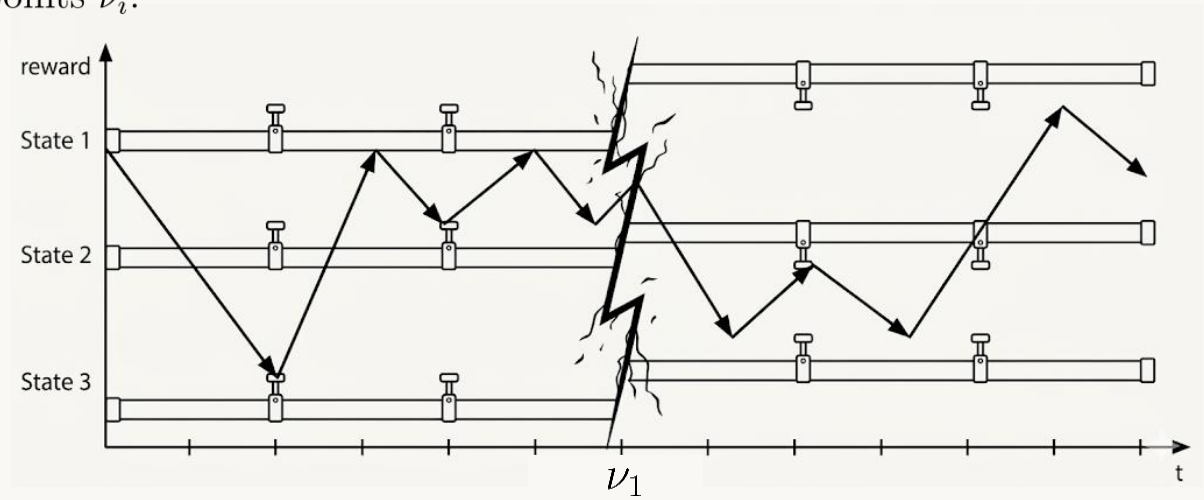
- **RMAB problem:** Similar to MAB problem, but evolves as a Markov chain regardless of whether it is selected
- **Objective:** Minimize Regret

- K: The total number of arms in the system.
- T: The total time horizon for the learning process.
- $\mathbf{P} = \{P_1, P_0\}$: The transition kernels of the Markovian arms.
- R: The reward distributions associated with each state.

Piecewise Stationary (PS)-RMAB Problem

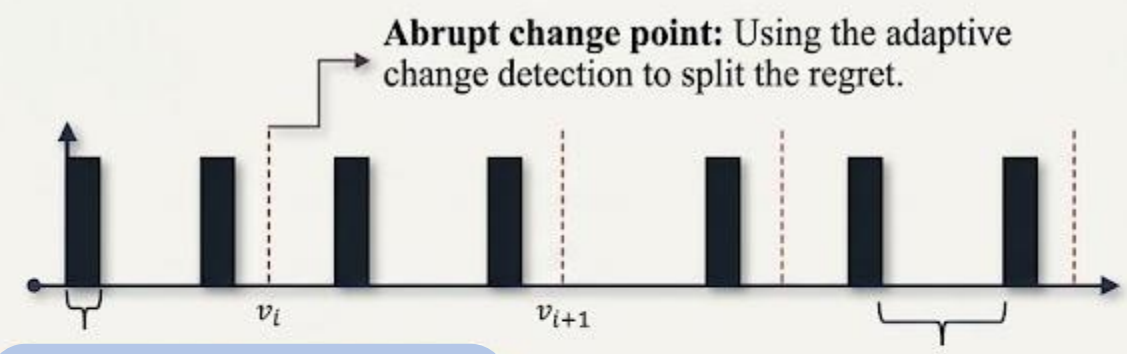
- **PS-RMAB:** RMAB but each arm's transition kernel and/or reward distribution may change once in a while.
  - Each arm k follows a Markovian transition kernel $\mathbf{P}_i$ and reward distribution $\mathbf{R}_i$ within segment i, and total have M segments.
  - Transition kernels or rewards change at unknown time points ν_i.

- **Challenge:**
  - The locations (ν_i) and the number (M) of change points are unknown.
  - Samples are not i.i.d.



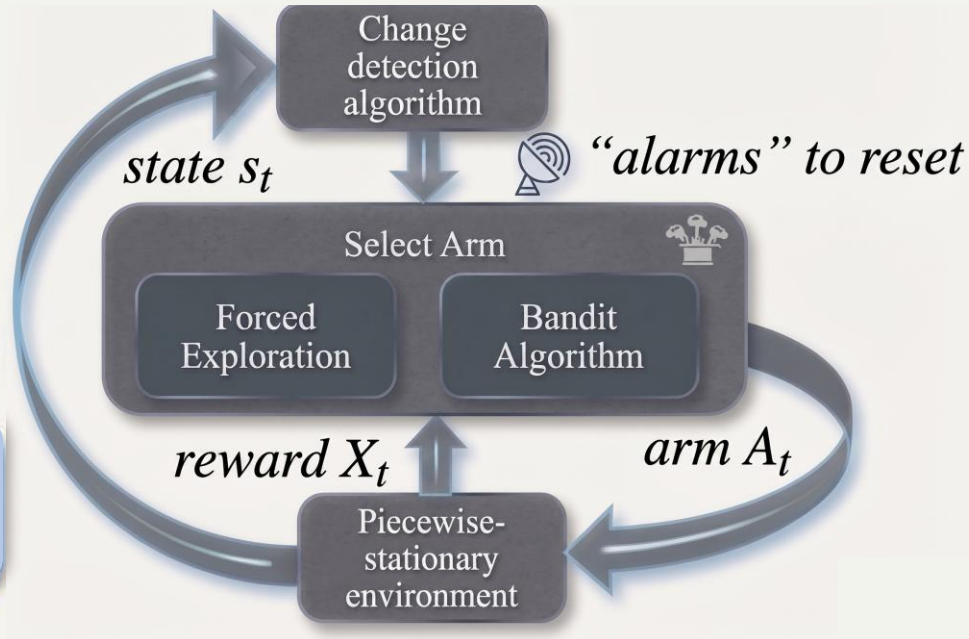
- **Excess Regret:** Measures the reward gap to a dynamic oracle that knows all ν_i .

Modularized Framework



Force exploration to ensure sufficient samples to detect change

Bandit base algorithm to deal with exploration vs exploitation

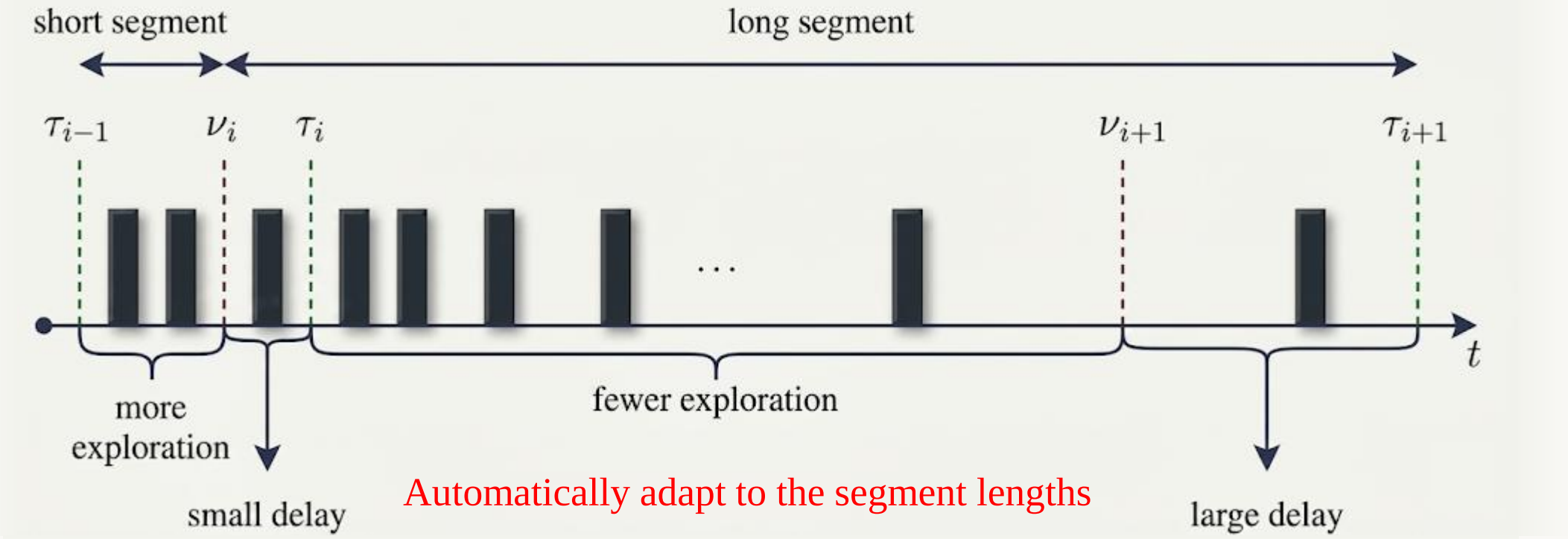


Challenge: How to balance the regret between the forced exploration and delay when M is unknown?

Proposed Forced Exploration: Diminishing Exploration

When segments are short:
Resets make us explore more frequently

When segments are long: Regrets caused by larger delay are compensated by fewer exploration



Automatically adapt to the segment lengths

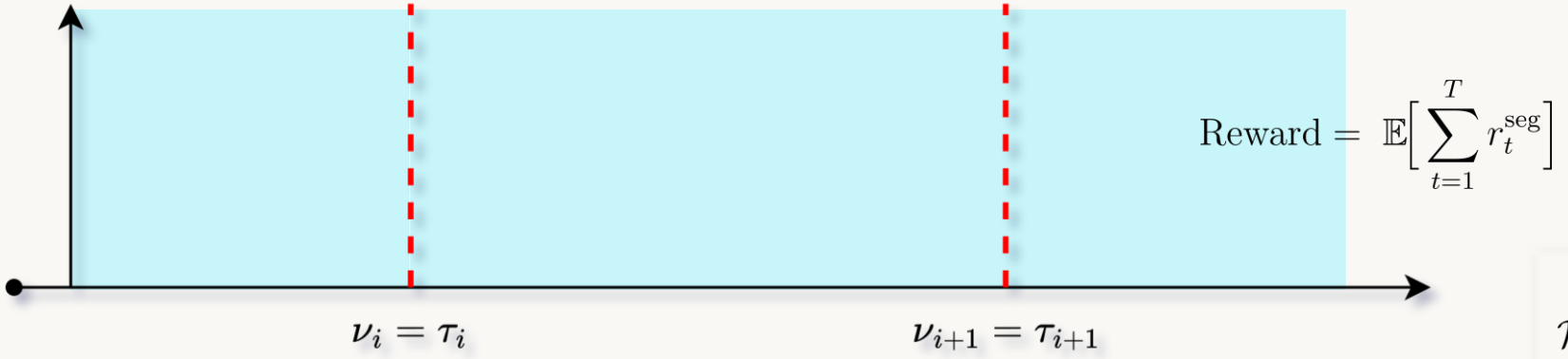
Theory

Base algorithm

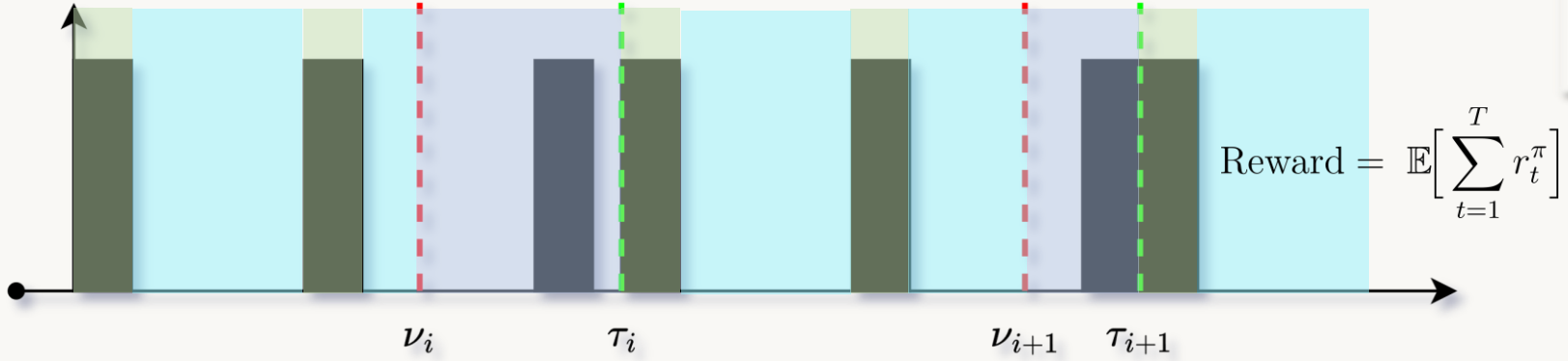
Force exploration

Delay

Ideal segmentation



Our policy π



$$\mathcal{R}_{\text{excess}}(T) = \mathbb{E}\Big[\sum_{t=1}^{T} r_t^{\text{seg}}\Big] - \mathbb{E}\Big[\sum_{t=1}^{T} r_t^{\pi}\Big]$$

$$= \text{Regret} + \text{Regret}$$

Theory

Corollary. *Combining diminishing exploration with resets with the base solver $\mathcal{B}$ and change detection $\mathcal{D}$, the excess regret is upper-bounded as*

$$\mathcal{R}_{\text{excess}}(T) \;\leq\; \mathcal{O}\Big(\sqrt{LKMT\log T}\Big)\,, \tag{1}$$

where L is the mixing time of the restless arms.

Known Lower Bounds:

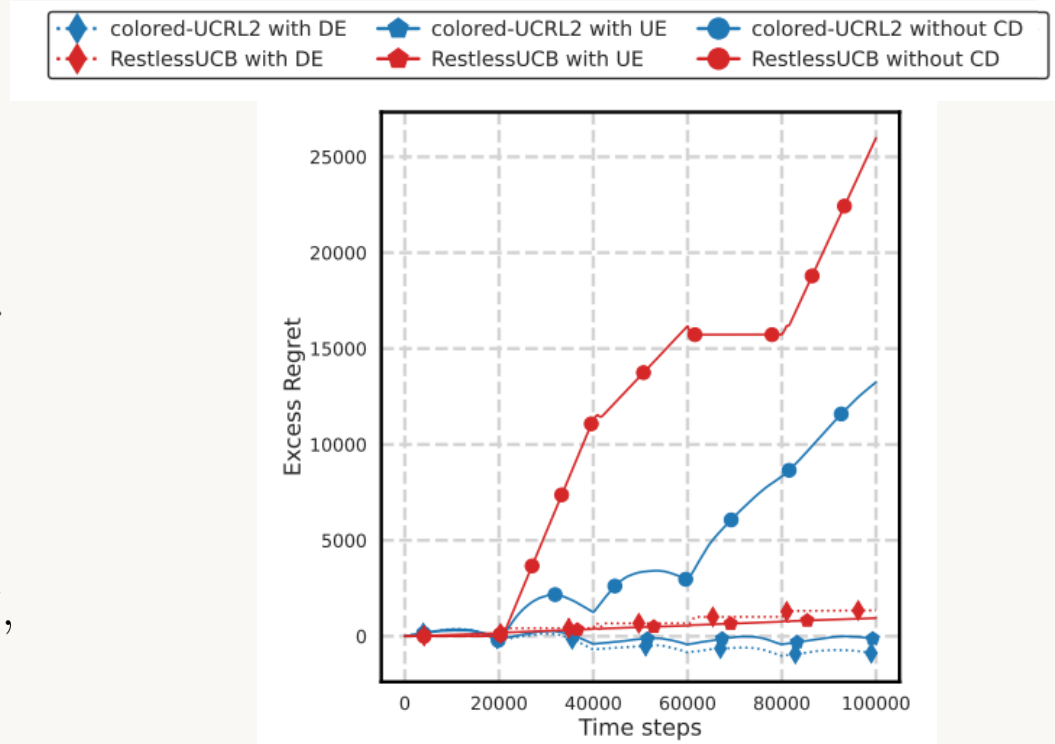
- Piecewise-stationary MAB: $\Omega(\sqrt{MKT})$ [Zhou et al., AAAI 2020]
- Stationary RMAB: $\Omega(\sqrt{T})$ [Ortner et al., 2012]

Our Upper Bound: $\mathcal{O}(\sqrt{LKMT\log T})$

Implication: By bridging the fundamental limits of both piecewise stationarity and restless dynamics, we reasonably infer that our modular framework achieves a nearly optimal regret bound.

Simulation

- **Base RMAB Solvers:** Evaluated using **RestlessUCB** and **colored-UCRL2**.
- **Exploration Strategies:**
  - **UE (Uniform Exploration):** Requires prior knowledge of segment count M. (Solid line)
  - **DE (Diminishing Exploration):** Does **not** require knowledge of M. (Dash line)
- **Performance Observation:** Without knowing M, DE-based frameworks achieve competitive or even superior regret performance compared to UE.



Excess Regret.