Globally Nonstationary Multi-Armed Bandits

Junpei Komiyama (NYU), Edouard Fouché (KIT), Junya Honda (Kyoto U)

July 29, 2021

Our manuscript is available on arXiv (and this slides)

Single page summary

- K-armed (multi-play) multi-armed bandit problem.
- We propose ADR-bandit algorithms that work well in both stationary and globally nonstationary environments.
  - Proposed algorithm: adaptive windows + stationary bandit.
  - Globally nonstationary environments = distributions of all the arms change in a coordinated manner.

this paper (→ ADR-bandit)

	Stationary	Abrupt	Gradual
Existing NS-MAB	$\tilde{O}(\sqrt{T})$	$\tilde{O}(\sqrt{T})$	$\tilde{O}(T^{1-d/3})$
ADR-bandit	$\boldsymbol{O(\log T/\Delta_{\min})}$	$\tilde{O}(\sqrt{T})$ (Under GC)	$\tilde{O}(T^{1-d/3})$ (Under GC)
Stationary MAB	$O(\log T/\Delta_{\min})$	$O(T)$	$O(T)$

regret bounds of the algorithms

Agenda

- **Introduction <- Next**
- Results on single stream: Total error of ADWIN algorithm
- Results on multi-armed bandits (MABs): Regret bound of ADR-bandit algorithm

Set up: nonstationary multiple-play MAB

- $\mu_{i,t}$: mean of arm $i \in \{1,2,\ldots,K\}$ at round t.
- At each round $t = 1,2,\ldots,T$, select $L < K$ arms and receives corresponding rewards $x_{i,t} \in [0,1]$
  - $L = 1$ (single-play MAB) as a special case
- Goal: maximize rewards by choosing L best arms.
- Partial observability (one only knows the reward of selected arms).
- Nonstationarity: Best arms change over time.
- Regret $= \sum_t (\max_{I:|I|=L} \mu_{i,t} - \sum_{i \in \text{selected}} \mu_{i,t})$

Literature

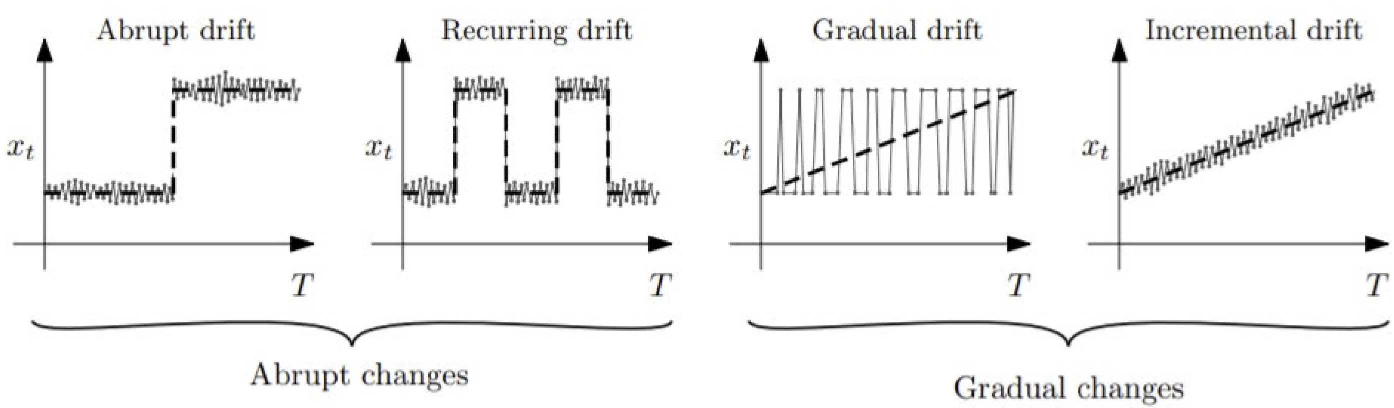
- Literature in non-stationary bandit problems includes:
  - Passive algorithms such as sliding windows (fixed size) [Garivier and Moulines 2008, Trovo+2020] and discounting [Kocsis and Szepesvari 2006], and
  - active algorithms involves change point detectors such as PH-test [Hartland+2007, Liu+2008], likelihood-ratio test [Besson+2020].
- Lower bound [Garivier and Moulines 2008]: Any non-stationary bandit algorithm (regardless of passive/active) for abrupt changes has at least $\Omega(\sqrt{T})$ forced exploration.
- Our algorithm is active and avoids forced exploration.

Agenda

- Introduction
- **Results on single stream: Total error of ADWIN algorithm <- Next**
- Results on multi-armed bandits: Regret bound of ADR-bandit algorithm

Data mining literature: Stream learning

- A stream is an unknown seq of means $(\mu_t)_t$.
  - = arm in our setting
- Stream is stationary if $\mu_t = \mu$ does not change over time, or
- Abruptly changing if $\mu_{t+1} \neq \mu_t$ at changepoints T_c, or
- Gradually changing if $|\mu_{t+1} - \mu_t| \leq T^{-d}$ for some constant $d \leq (0,1)$.



$\mu_{i,t}$ changes quickly at changepoints

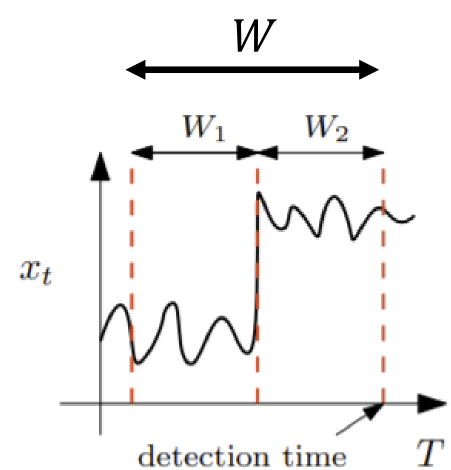
$\mu_{i,t}$ changes slowly at changepoints

Adaptive windowing (ADWIN)

[Bifet and Gavalda 2007]

- Q: If we observe all the rewards $x_1, x_2, x_3, \ldots, x_{t-1}$, how accurate one can estimate mean μ_t of the current round?
- ADWIN: dynamic control of window W
  - $\hat{\mu}_w$: Empirical mean based on time steps in $W = \{s, s+1, s+2, \ldots, t-1\}$.
- Shrinks W when it detects significant gap.

Idea:
Cuts W_1 if it detects a significant gap between the two emp.means.
- Checks every split $W = W_1 \cup W_2$



Analysis of adaptive windowing

- [BG2007] bounds false positive and negative (fp/fn) rates of ADWIN for given t.
  - Cons: fp/fn rates do not directly help bandit analysis.
- Instead, we bound total error: $\mathrm{Err(T)} = \sum_{t=1}^{T} |\hat{\mu}_W - \mu_t|$

Results in abruptly changing streams:

- Thm 8 (**abrupt**): Total error of ADWIN in an abruptly changing stream with M changes is $\tilde{\mathrm{O}}(\sqrt{MT})$.
- This results is strong, there is NO requirement on the change point size, distance, and the algorithm does NOT need to know M.

Analysis of adaptive windowing

- Thm 10 (**gradual**): Total error of ADWIN in a gradually changing stream is $\tilde{O}(T^{1-d/3})$.
  - Useful lemma (Lemma 27):

    For any $N \geq |W|, |\mu_t - \mu_W| \leq 3T^{-d}N + \tilde{O}\left(\frac{1}{\sqrt{N}}\right)$.
  - $N = T^{2d/3}$ gives the bound of Thm 10.

change speed $= T^{-d}$

Agenda

- Introduction
- Results on single stream: Total error of ADWIN algorithm
- **Results on multi-armed bandits: Regret bound of ADR-bandit algorithm <- Next**

Scaling bandits
[Fouché, Komiyama, Böhm, KDD2019]

- Idea of adaptive window + Thompson sampling was introduced [FKB2019].
  - No analysis for nonstationary bandits, though it empirically performed very well.
- Unlike [FKB2019], this paper gives the regret bound for such algorithms.

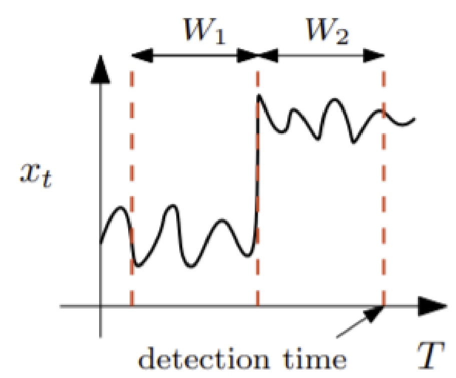
Proposed algorithm: ADR-bandits

- Adaptive resetting (ADR) bandit consists of

1. Base bandit algorithm (ex: Thompson sampling, KL-UCB).
2. Change point detectors (ADRs) for each of K arms.

- At each round, select arms by using the base bandit algorithm. When one of the ADRs detects a change, reset the entire algorithm.

Idea:
Reset entire alg if it detects a significant gap between the two emp.means.
Checks every split $W = W_1 \cup W_2$ and every arm.



Setting up analysis: Characterize assumptions on base bandit algorithm

The properties of base bandit algorithms:

1. (distribution-dependent regret) In a stationary env, it has an $O(\log \mathrm{T}/\Delta)$ regret bound.

2. (drift-tolerant regret) In a nonstationary env, it has an $\tilde{O}(\sqrt{KT} + \epsilon(T))$ regret bound

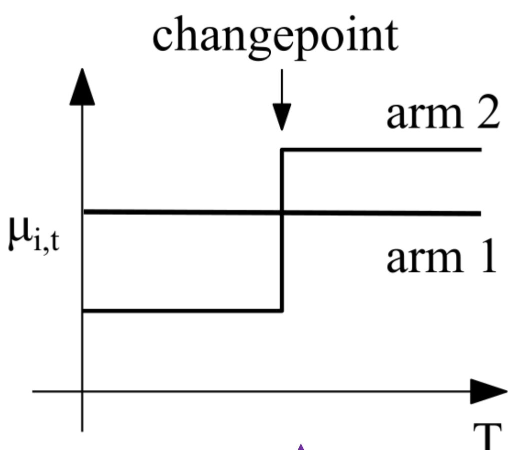
    - $\epsilon(t) =$ drift of $\mu_{i,t}$.

3. (monitoring consistency) At least one arms is selected consistently.

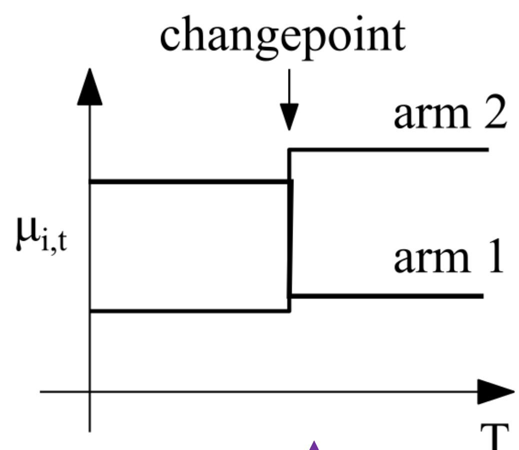
- We explicitly show an algorithm that has these properties (E-UCB, Algorithm 5)

Assumption on the streams: Global changes

- Our algorithm works on stationary and globally changing environments.
- Global changes: All the arms changes in the coordinated manner.
- Similar idea [Mukherjee and Maillard 2019] for abrupt case.



NON-global abrupt change



global abrupt change

Main results: Regret bounds

- Under the assumptions on the previous pages:

Stationary case:

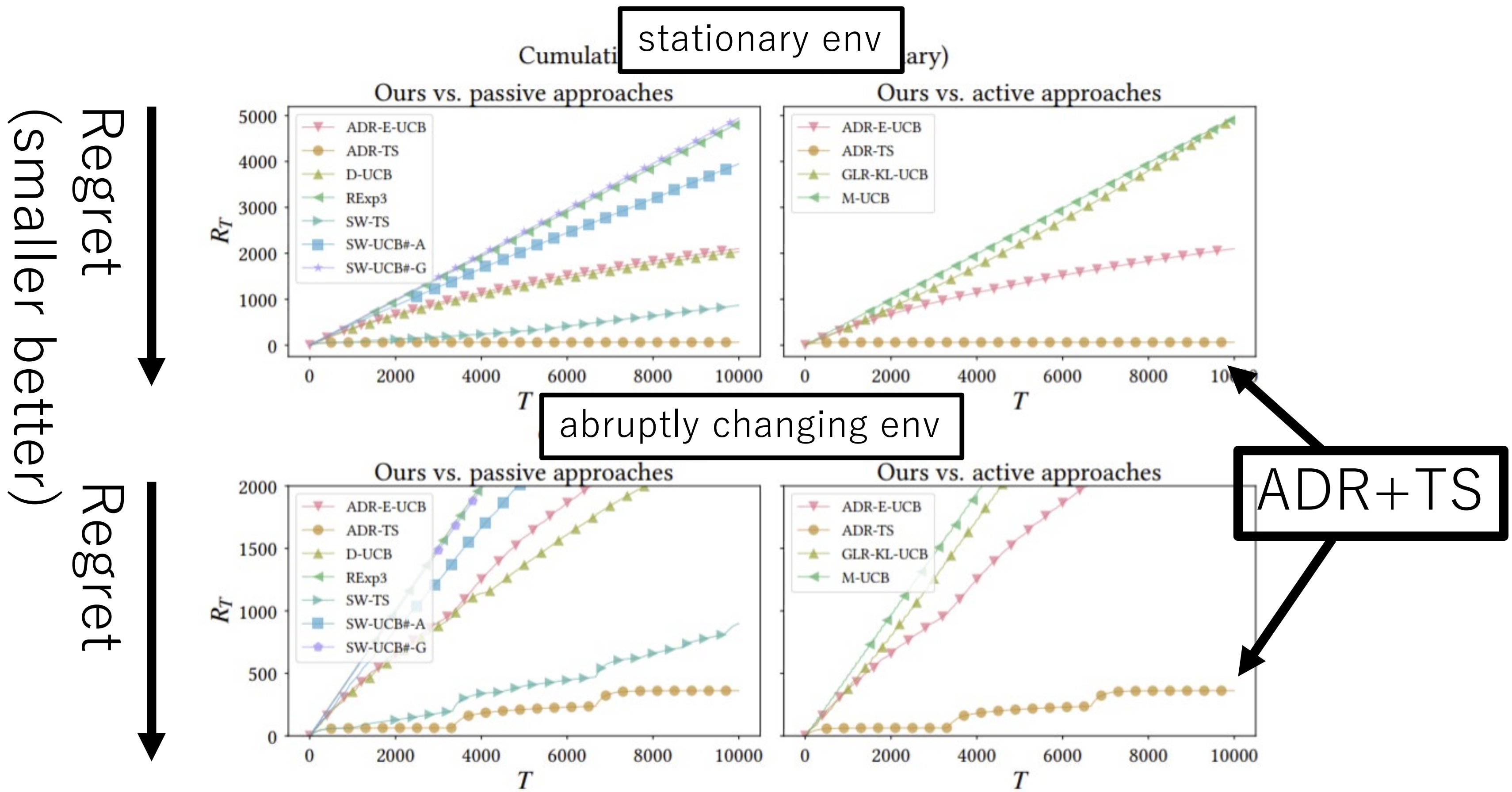
- Theorem 20: In a stationary env, the regret of ADR-bandit is $O\left(\frac{\log T}{\Delta}\right)$.

disclaimer:
requires detectability condition

Non-stationary cases:

- Theorem 22: In an env with globally abrupt changes, the regret of ADR-bandit is $O(\sqrt{MKT})$.
- Theorem 24: In an env with globally gradual changes, the regret of ADR-bandit is $O\left((\sqrt{LK})\mathrm{T}^{1-\frac{\mathrm{d}}{3}}\right)$.

Experimental results



- More results on the paper

Single page summary (recap)

- K-armed (multi-play) multi-armed bandit problem.
- We propose ADR-bandit algorithms that work well in both stationary and globally nonstationary environments.
  - Adaptive windows + stationary bandit algs.
  - Globally nonstationary environments = distributions of all the arms change in a coordinated manner.

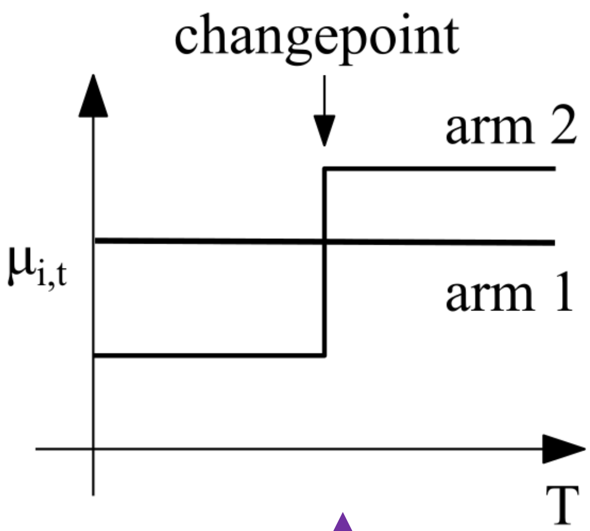
this paper

	Stationary	Abrupt	Gradual
Existing NS-MAB	$\tilde{O}(\sqrt{T})$	$\tilde{O}(\sqrt{T})$	$\tilde{O}(T^{1-d/3})$
ADR-bandit	$\boldsymbol{O(\log T/\Delta_{\min})}$	$\tilde{O}(\sqrt{T})$ (Under GC)	$\tilde{O}(T^{1-d/3})$ (Under GC)
Stationary MAB	$O(\log T/\Delta_{\min})$	$O(T)$	$O(T)$

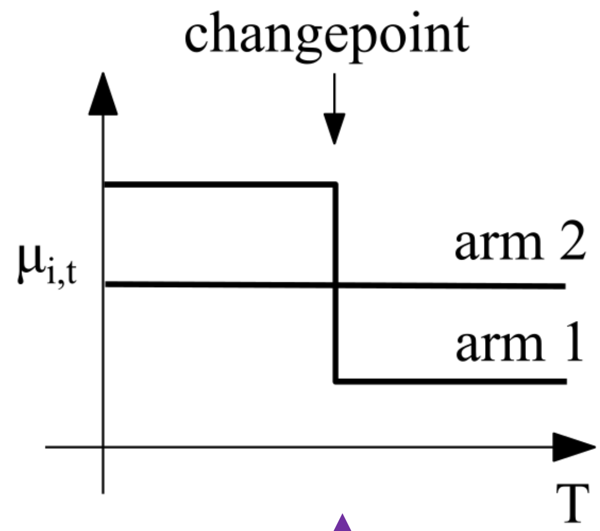
regret bounds of the algorithms

Future works (last slide)

- Further characterization of non-stationary bandits that extend stationary bandits: There are some NON-global changes that stationary bandit algorithm can deal with.



hard non-global abrupt change



easy non-global abrupt change