Online Learning in Rested and Restless Bandits
Workshop on restless bandits, index policies and applications in reinforcement learning

Cem Tekin
Bilkent University

November 20, 2023
Background

- I worked on rested and restless bandits during my PhD (2008-2013)
  - U Michigan
  - Demos Teneketsiz, Mingyan Liu, Ambuj Tewari..
- The talk is mainly about my PhD work:
- Remarkable progress in the field since then
$k$-armed i.i.d. bandit problem

- $k$ arms with fixed and unknown reward distributions (frequentist setting)

\[ \nu = \nu_1 \times \nu_2 \times \ldots \times \nu_k \]

- Unknown expected arm rewards:

\[ \mu_1 \geq \mu_2 \geq \ldots \geq \mu_k \]

- $n$ rounds of sequential interaction. At round $t$:
  - Learner plays arm $A_t \in [k]$
  - Learner observes noisy reward $X_{A_t}(t) \sim \nu_{A_t}$
Regret for i.i.d. bandit problem

- Goal: For a given horizon $n$

\[
\text{Maximize } \mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t}(t) \right]
\]

- Optimal policy when $\mu_1, \ldots, \mu_k$ are known:

  Always choose arm 1

- Minimize frequentist regret:

\[
R(n) = n\mu_1 - \mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t}(t) \right]
\]
UCB algorithm for i.i.d. bandit [Auer et al. 2002]

- Any meaningful policy samples suboptimal arm \( i \) \( \Omega(\log n) \) times [Lai & Robbins, 1985]

**Algorithm 1: UCB**

```plaintext
for \( t = 1, 2, \ldots \) do
  1. Compute UCB indices: \( \text{UCB}_i(t) = \hat{\mu}_i(t) + \sqrt{\frac{\alpha \log(t-1)}{T_i(t)}} \)
  2. Play arm \( A_t = \arg\max_{i \in [k]} \text{UCB}_i(t) \)
end
```

\( \alpha \): exploration constant; \( T_i(t) \): number of plays of arm \( i \) before \( t \); \( \hat{\mu}_i(t) \): sample mean reward of arm \( i \)

**Properties of UCB:**
- Achieves \( O(\log n) \) instance-dependent regret (order-optimal)
- Anytime (no need to know \( n \))
- Compute & memory efficient
Any meaningful policy samples suboptimal arm \( i \) \( \Omega(\log n) \) times [Lai & Robbins, 1985]

**Algorithm 2: UCB**

```plaintext
for \( t = 1, 2, \ldots \) do

1. Compute UCB indices: \( \text{UCB}_i(t) = \hat{\mu}_i(t) + \sqrt{\frac{\alpha \log(t-1)}{T_i(t)}} \)

2. Play arm \( A_t = \arg \max_{i \in [k]} \text{UCB}_i(t) \)

end
```

\( \alpha \): exploration constant; \( T_i(t) \): number of plays of arm \( i \) before \( t \); \( \hat{\mu}_i(t) \): sample mean reward of arm \( i \)

**Properties of UCB:**
- Achieves \( O(\log n) \) instance-dependent regret (order-optimal)
- Anytime (no need to know \( n \))
- Compute & memory efficient
Any meaningful policy samples suboptimal arm $i \Omega(\log n)$ times [Lai & Robbins, 1985]

Algorithm 3: UCB

\begin{align*}
\text{for } t = 1, 2, \ldots \text{ do} \\
1. \text{Compute UCB indices: } UCB_i(t) &= \hat{\mu}_i(t) + \sqrt{\frac{\alpha \log(t-1)}{T_i(t)}} \\
2. \text{Play arm } A_t = \arg \max_{i \in [k]} UCB_i(t) \\
\end{align*}
\[
\alpha: \text{ exploration constant; } T_i(t): \text{ number of plays of arm } i \text{ before } t; \hat{\mu}_i(t): \text{ sample mean reward of arm } i
\]

Properties of UCB:
- Achieves $O(\log n)$ instance-dependent regret (order-optimal)
- Anytime (no need to know $n$)
- Compute & memory efficient
**k**-armed rested bandit problem

**Arm** \(i\):
- Finite state space \(S_i\)
- Reward = state (noiseless observations)
- When not played, state remains **frozen**
- When played, state transitions according to **unknown** \(P_i\)
- When played, an **irreducible, aperiodic** Markov chain

![Diagram of arm transitions](image-url)
Regret for rested bandit problem

- Maximize $\mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t}(t) \middle| x_0 \right]$ over horizon $n$?

- Optimal policy when $P_i$s are known? **non-stationary, intractable**

- Minimize the following regret?

$$R(n) = \mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t^*}(t) \middle| x_0 \right] - \mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t}(t) \middle| x_0 \right]$$

Too ambitious!
Regret for rested bandit problem

- Maximize $\mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t}(t) \middle| x_0 \right]$ over horizon $n$?
- Optimal policy when $P_i$s are known? **non-stationary, intractable**
- Minimize the following regret?

$$R(n) = \mathbb{E} \left[ \sum_{t=1}^{n} X_{A^*_t}(t) \middle| x_0 \right] - \mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t}(t) \middle| x_0 \right]$$

Too ambitious!
Regret for rested bandit problem

- Maximize $\mathbb{E}\left[\sum_{t=1}^{n} X_{A_t}(t) \bigg| x_0\right]$ over horizon $n$?
- Optimal policy when $P_i$s are known? non-stationary, intractable
- Minimize the following regret?

$$R(n) = \mathbb{E}\left[\sum_{t=1}^{n} X_{A^*_t}(t) \bigg| x_0\right] - \mathbb{E}\left[\sum_{t=1}^{n} X_{A_t}(t) \bigg| x_0\right]$$

optimal policy

learner's policy

Too ambitious!
Regret for rested bandit problem

- Maximize $\mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t}(t) \bigg| x_0 \right]$ over horizon $n$?
- Optimal policy when $P_i$s are known? **non-stationary, intractable**
- Minimize the following regret?

$$R(n) = \mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t^*}(t) \bigg| x_0 \right] - \mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t}(t) \bigg| x_0 \right]$$

Too ambitious!
Let’s try something else

- For $0 < \beta < 1$, Maximize $\mathbb{E} \left[ \sum_{t=1}^{\infty} \beta^{t-1} X_{A_t}(t) \bigg| x_0 \right]$?

- Optimal policy when $P_1, \ldots, P_k$ are known? **Gittins index policy**

- Minimize the following regret?

$$R(n) = \mathbb{E} \left[ \sum_{t=1}^{n} \beta^{t-1} X_{A^*_t}(t) \bigg| x_0 \right] - \mathbb{E} \left[ \sum_{t=1}^{n} \beta^{t-1} X_{A_t}(t) \bigg| x_0 \right]$$

Not enough time to learn!
Alternative regret for rested bandit problem

Let’s try something else

- For $0 < \beta < 1$, Maximize $\mathbb{E}\left[\sum_{t=1}^{\infty} \beta^{t-1} X_{A_t}(t) \mid x_0\right]$?
- Optimal policy when $P_1, \ldots, P_k$ are known? Gittins index policy
- Minimize the following regret?

$$R(n) = \mathbb{E}\left[\sum_{t=1}^{n} \beta^{t-1} X_{A^*_t}(t) \mid x_0\right] - \mathbb{E}\left[\sum_{t=1}^{n} \beta^{t-1} X_{A_t}(t) \mid x_0\right]$$

Gittins index policy

learner’s policy

Not enough time to learn!
Alternative regret for rested bandit problem

Let’s try something else

- For $0 < \beta < 1$, Maximize $\mathbb{E} \left[ \sum_{t=1}^{\infty} \beta^{t-1} X_{A_t}(t) \bigg| x_0 \right]$?
- Optimal policy when $P_1, \ldots, P_k$ are known? **Gittins index policy**
- Minimize the following regret?

$$R(n) = \mathbb{E} \left[ \sum_{t=1}^{n} \beta^{t-1} X_{A_t}(t) \bigg| x_0 \right] - \mathbb{E} \left[ \sum_{t=1}^{n} \beta^{t-1} X_{A_t}(t) \bigg| x_0 \right]_{\text{Gittins index policy}} - \mathbb{E} \left[ \sum_{t=1}^{n} \beta^{t-1} X_{A_t}(t) \bigg| x_0 \right]_{\text{learner’s policy}}$$

Not enough time to learn!
Alternative regret for rested bandit problem

Let’s try something else

- For $0 < \beta < 1$, Maximize $\mathbb{E} \left[ \sum_{t=1}^{\infty} \beta^{t-1} X_{A_t}(t) \bigg| x_0 \right]$?

- Optimal policy when $P_1, \ldots, P_k$ are known? Gittins index policy

- Minimize the following regret?

$$R(n) = \mathbb{E} \left[ \sum_{t=1}^{n} \beta^{t-1} X_{A^*_t}(t) \bigg| x_0 \right] - \mathbb{E} \left[ \sum_{t=1}^{n} \beta^{t-1} X_{A_t}(t) \bigg| x_0 \right]$$

Gittins index policy learner’s policy

Not enough time to learn!
Weak regret for rested bandit problem

Let's try something simpler

- Let $\{\pi_i(x)\}_{x \in S_i}$ represent the unique stationary distribution of arm $i$

  $$\mu_i := \lim_{n \to \infty} \frac{1}{n} \mathbb{E} \left[ \sum_{t=1}^{n} X_i(t) \mid x_0 \right] = \sum_{x \in S_i} x \pi_i(x)$$

- Assume $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_k$

- Optimal policy when $P_1, \ldots, P_k$ are known? Since arms are rested and ergodic, as $n \to \infty$, for the optimal policy $A^*_t = 1$ for $t$ large.

- Minimize the weak regret

  $$R_w(n) = \underbrace{n \mu_1}_{\text{proxy for the opt}} - \mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t}(t) \mid x_0 \right]$$

  learner’s policy
Weak regret for rested bandit problem

Let’s try something simpler

- Let \( \{\pi_i(x)\}_{x \in S_i} \) represent the unique stationary distribution of arm \( i \)

\[
\mu_i := \lim_{n \to \infty} \frac{1}{n} \mathbb{E} \left[ \sum_{t=1}^{n} X_i(t) \right | x_0] = \sum_{x \in S_i} x \pi_i(x)
\]

- Assume \( \mu_1 \geq \mu_2 \geq \ldots \geq \mu_k \)

- Optimal policy when \( P_1, \ldots, P_k \) are known? Since arms are rested and ergodic, as \( n \to \infty \), for the optimal policy \( A^*_t = 1 \) for \( t \) large.

- Minimize the weak regret

\[
R_w(n) = n \mu_1 - \mathbb{E} \left[ \sum_{t=1}^{n} X_{A^*_t}(t) \right | x_0] = n \mu_1 - \mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t}(t) \right | x_0]
\]
Weak regret for rested bandit problem

Let's try something simpler

- Let \( \{\pi_i(x)\}_{x \in S_i} \) represent the unique stationary distribution of arm \( i \)

\[
\mu_i := \lim_{n \to \infty} \frac{1}{n} \mathbb{E} \left[ \sum_{t=1}^{n} X_i(t) \bigg| x_0 \right] = \sum_{x \in S_i} x \pi_i(x)
\]

- Assume \( \mu_1 \geq \mu_2 \geq \ldots \geq \mu_k \)

- Optimal policy when \( P_1, \ldots, P_k \) are known? Since arms are rested and ergodic, as \( n \to \infty \), for the optimal policy \( A_t^* = 1 \) for \( t \) large.

- Minimize the weak regret

\[
R_w(n) = \underbrace{n \mu_1}_{\text{proxy for the opt}} - \mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t}(t) \bigg| x_0 \right]
\]
Algorithm 4: UCB-rested

for $t = 1, 2, \ldots$ do

1. Compute UCB indices: $UCB_i(t) = \hat{\mu}_i(t) + \sqrt{\frac{\alpha \log(t-1)}{T_i(t)}}$

2. Play arm $A_t = \arg\max_{i \in [k]} UCB_i(t)$

end

$\alpha$: exploration constant; $T_i(t)$: number of plays of arm $i$ before $t$; $\hat{\mu}_i(t)$: sample mean reward of arm $i$

**Difference from i.i.d. UCB?**
Choice of $\alpha$ that yields $O(\log n)$ instance-dependent regret depends on state space cardinality and eigenvalue gap of transition matrices.
Instance-dependent regret bound

Conditions for the regret bound:
- All arms are finite-state, irreducible, aperiodic Markov chains with $P_i$s having irreducible multiplicative symmetrizations (MS)
- For any state $x$ of any arm, $0 < x < 1$
- $\epsilon_i$: eigenvalue gap of MS of $P_i$
- $\epsilon_{\text{min}} = \min_i \epsilon_i$
- $S_{\text{max}} = \max_{i \in [k]} |S_i|$

Theorem

When UCB is run with $\alpha = O\left(\frac{S_{\text{max}}^2}{\epsilon_{\text{min}}}\right)$, we have

$$R_w(n) = O\left(\frac{S_{\text{max}}^2}{\epsilon_{\text{min}}} \sum_{i: \mu_i < \mu_1} \frac{\log n}{\mu_1 - \mu_i}\right)$$
Learning policies that compete with Gittins index?

A recent paper by [Gast et al. 2022]¹

- Discount factor $\beta < 1$
- Episodic setting with $n$ episodes and geometrically distributed episode lengths
- Computationally tractable algorithms with Bayesian strong regret bound of $O(S_{\text{max}} \sqrt{nK})$

---

¹ Learning algorithms for Markovian bandits: Is posterior sampling more scalable than optimism?, TMLR, 2022
k-armed restless bandit problem

Arm $i$:

- All assumptions same as rested **except**:
  - When not played, state transitions *arbitrarily*
  - Only state of the played arm is observed
An application of restless bandit problem

Opportunistic spectrum access or cognitive radio

[Diagram of restless bandit problem with multiple channels and transitions between good and bad states.]
Regret for restless bandit problem

- Maximize \( \mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t}(t) \right] \) over horizon \( n \)?
- Optimal policy when \( P_i \)'s are known? non-stationary, intractable
- Minimize the following regret?

\[
R(n) = \mathbb{E} \left[ \sum_{t=1}^{n} X_{A^*_t}(t) \right] - \mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t}(t) \right]
\]

Too ambitious!
Regret for restless bandit problem

- Maximize $\mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t}(t) \bigg| x_0 \right]$ over horizon $n$?
- Optimal policy when $P_i$s are known? **non-stationary, intractable**
- Minimize the following regret?

$$R(n) = \mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t^*}(t) \bigg| x_0 \right] - \mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t}(t) \bigg| x_0 \right]$$

too ambitious!
Regret for restless bandit problem

- Maximize $\mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t}(t) \bigg| x_0 \right]$ over horizon $n$?
- Optimal policy when $P_i$s are known? **non-stationary, intractable**
- Minimize the following regret?

$$R(n) = \mathbb{E} \left[ \sum_{t=1}^{n} X_{A^*_t}(t) \bigg| x_0 \right] - \mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t}(t) \bigg| x_0 \right]$$

Too ambitious!
Regret for restless bandit problem

- Maximize $\mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t}(t) \mid x_0 \right]$ over horizon $n$?
- Optimal policy when $P_i$s are known? **non-stationary, intractable**
- Minimize the following regret?

$$R(n) = \mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t^*}(t) \mid x_0 \right] - \mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t}(t) \mid x_0 \right]$$

Too ambitious!
Alternative regret for restless bandit problem

- Recall that for i.i.d. and rested bandits our “weak” benchmark was $n\mu_1$.
- We seek to minimize the weak regret:

  $$R_w(n) = n\mu_1 - \mathbb{E}\left[\sum_{t=1}^{n} X_{A_t}(t) \mid x_0\right]$$

Why?
- Tradeoff between performance and complexity
- Scalability for compute and memory constrained, battery dependent devices
- In line with satisficing principle of Herbert Simon
  “Decision makers can satisfice either by finding optimum solutions for a simplified world or by satisfactory solutions for a more realistic world”
Alternative regret for restless bandit problem

- Recall that for i.i.d. and rested bandits our “weak” benchmark was $n\mu_1$.
- We seek to minimize the weak regret:

$$R_w(n) = n\mu_1 - \mathbb{E} \left[ \sum_{t=1}^{n} X_{A_t}(t) \bigg| x_0 \right]$$

Why?

- Tradeoff between performance and complexity
- Scalability for compute and memory constrained, battery dependent devices
- In line with satisficing principle of Herbert Simon
  
  “Decision makers can satisfice either by finding optimum solutions for a simplified world or by satisfactory solutions for a more realistic world”
A connection between rested and restless bandits

- Let $\tau_i(m)$ represent the time index of $m$th play of arm $i$
- For a rested arm $\hat{\mu}_i(t) = \frac{X_i(\tau_i(1)) + \ldots + X_i(\tau_i(t))}{T_i(t)} \to \mu_i$
- For a restless arm $\hat{\mu}_i(t) \not\to \mu_i$ since $X_i(\tau_i(1)), \ldots, X_i(\tau_i(t))$ do not form a continuous sample path for “active” Markov chain of arm $i$

**Design an algorithm that stitches together discontinuous segments of observations from a restless arm to form a rested arm with the same $P_i$ as the restless arm**
The regenerative cycle algorithm (RCA) [Tekin & Liu, 2012]

- An arm is played in blocks till a full regenerative cycle is observed (starting in some state $\gamma_i$ and ending in $\gamma_i$).
- Since arm selections are interleaved, observations from an arm are carefully stitched together to mimic a rested arm.
  - Block $B = [SB1, SB2, SB3]$
  - $SB1$: Play till $\gamma_i$ is hit
  - $SB2, SB3$: Play till $\gamma_i$ is hit again

```
\begin{align*}
\cdots & \gamma_i \cdots 1 \gamma_i \cdots j \cdots \gamma_i \cdots 2 \gamma_i \cdots \\
SB1 & \quad SB2 & \quad SB3 & \quad SB1 & \quad SB2 & \quad SB3 & \quad SB2 & \quad SB3
\end{align*}
```

- \[ \cdots \text{play } i \cdots \text{play } j \cdots \text{play } i \cdots \]

- \[ \cdots \cdots \cdots \cdots \gamma_i^* \gamma_i^* \gamma_i^* \gamma_i^* \gamma_i^* \cdots \cdots \cdots \cdots \]

- \[ \cdots \cdots \cdots \cdots \gamma_i \cdots 3 \gamma_i \cdots \cdots \cdots \cdots \]

- \[ \cdots \cdots \cdots \cdots \text{SB1} \quad \text{SB2} & \quad \text{SB3} \]


C. Tekin
OL in Rested and Restless Bandits
November 20, 2023
The regenerative cycle algorithm (RCA)

When we stitch together SB2s of arm $i$:

A continuous sample path from $P_i$ (rested UCB analysis apply)

Moreover, blocks are i.i.d. by the regenerative cycle theorem [Brémaud Thm. 7.4.]
Algorithm 5: RCA-i.i.d.

At the end of $b$th block:
1. Compute UCB indices: $UCB_i(b+1) = \frac{Y_{i,2}(b)}{N_{i,2}(b)} + \sqrt{\frac{\alpha \log b}{B_i(b)}}$
2. Play arm $A_{b+1} = \arg \max_{i \in [k]} UCB_i(b+1)$

- $B_i(b)$: number of completed blocks of arm $i$ so far
- $N_{i,2}(b)$: number of rounds spent in SB2 of arm $i$ so far
- $Y_{i,2}(b)$: cumulative reward from SB2 of arm $i$ so far

$$i = \arg \max_{a \in [k]} UCB_a(b+1)$$
RCA based on i.i.d. property of the regenerative cycles

**Theorem**

When $0 < x < 1$ for all $x \in S_i$ and $\alpha = 2$, the weak regret of RCA is

$$R_w(n) = O \left( \sum_{i: \mu_i < \mu_1} \frac{\log n}{(\mu_1 - \mu_i)^2} \left[ \underbrace{(\mu_1 - \mu_i)}_{A} \mathbb{E}_{\gamma_i}[SB2] + \underbrace{\mathbb{E}_{worst}[SB1]}_{C} \right] \right)$$

- A: Number of blocks (up to $n$) where suboptimal arm $i$ selected
- B: Expected regret in regenerative cycle of arm $i$ is block
- C: Expected regret from SB1 before hitting $\gamma_i$ in arm $i$ is block

Plug-in your favorite i.i.d. bandit algorithm. RCA should work.
RCA based on continuous sample path property

- $\gamma_i$ can be updated online to form the longest continuous sample path from arm $i$
  - SB2s of arm $i$ are no longer i.i.d.
  - Rested analysis over SB2s still apply

Use cases:
- Arrange $\gamma_i$s to minimize “wasted” observations in SB1s
- Can update indices when the task assigned to the arm completes
- For rested-based analysis, UCB **exploration constant** $\alpha$ requires knowledge of minimum eigenvalue gap (an instance-dependent quantity)
  - Grow $\alpha$ slowly over time $\alpha(n) \to \infty$
- Play $M$ arms each time
  - At each round arms with highest $M$ indices are played
- Rested bandits: analysis straightforwardly extends
- Restless bandits: need to account for random block lengths
For rested-based analysis, UCB exploration constant $\alpha$ requires knowledge of minimum eigenvalue gap (an instance-dependent quantity)
  - Grow $\alpha$ slowly over time $\alpha(n) \to \infty$

Play $M$ arms each time
  - At each round arms with highest $M$ indices are played
  - Rested bandits: analysis straightforwardly extends
  - Restless bandits: need to account for random block lengths
Other approaches for log weak regret in restless bandits

- Deterministic sequencing of exploration and exploitation (DSEE) [Liu et al. 2013]
  - Exploration & exploitation blocks are separate
  - All arms explored same amount of time
  - Geometrically increasing block lengths to wash away transient effects
  \[ R_W(n) = O \left( \frac{\log n}{\epsilon_{\text{min}}(\mu_1 - \mu_2)^2} \right) \]

- Adaptive sequencing rule (ASR) [Gafni & Cohen, 2021]
  - Exploration & exploitation blocks are separate
  - Geometrically increasing block lengths to wash away transient effects
  - Tries to explore arm \( i \) about \( O \left( \frac{\log n}{(\mu_1 - \mu_i)^2} \right) \) times by estimating the gap
  - Uses RCA within exploration blocks to form accurate estimates of arm means
  \[ R_W(n) = O \left( \frac{\log n}{\epsilon_{\text{min}}(\mu_1 - \mu_2)} \right) \] with tuned parameters
Other approaches for log weak regret in restless bandits

- Deterministic sequencing of exploration and exploitation (DSEE) [Liu et al. 2013]
  - Exploration & exploitation blocks are separate
  - All arms explored same amount of time
  - Geometrically increasing block lengths to wash away transient effects
  - \( R_W(n) = O \left( \frac{\log n}{\epsilon_{\text{min}}(\mu_1 - \mu_2)^2} \right) \)

- Adaptive sequencing rule (ASR) [Gafni & Cohen, 2021]
  - Exploration & exploitation blocks are separate
  - Geometrically increasing block lengths to wash away transient effects
  - Tries to explore arm \( i \) about \( O(\frac{\log n}{(\mu_1 - \mu_i)^2}) \) times by estimating the gap
  - Uses RCA within exploration blocks to form accurate estimates of arm means
  - \( R_W(n) = O \left( \frac{\log n}{\epsilon_{\text{min}}(\mu_1 - \mu_2)} \right) \) with tuned parameters
Learning policies that compete with Whittle index?

A recent preprint by [Akbarzadeh & Mahajan, 2023]²

- Undiscounted $\beta = 1$
- Uncontrolled transitions according to $P_i$ independent of active or passive
- Bayesian strong regret bound of $\tilde{O}(KS_{\max}\sqrt{n})$

² On learning Whittle index policy for restless bandits with scalable regret
Research directions

- Understanding dependence of instance-dependent weak regret on $S_{\text{max}}$
- Improving gap-dependence of weak regret in restless bandits
- Frequentist analysis w.r.t. other benchmarks (e.g., Gittins, Whittle)

THANK YOU!
Research directions

- Understanding dependence of instance-dependent weak regret on $S_{\text{max}}$
- Improving gap-dependence of weak regret in restless bandits
- Frequentist analysis w.r.t. other benchmarks (e.g., Gittins, Whittle)

THANK YOU!