Phase Transitions and Cyclic Phenomena in Bandits with Switching Constraints

David Simchi-Levi Yunzong Xu

Institute for Data, Systems and Society

MIT

Talk Outline

- **Motivation**
  - Practical & Theoretical Motivation
- **MAB with Unit Switching Cost**
  - Limited Number of Switches
- **Extensions**
  - General Switching Costs

Stochastic Multi-Armed Bandit (MAB)

- Finite actions/arms: $1, \ldots, k$
- A total of T rounds, $T \gg k$
- The unknown reward distribution of action i is D_i; expected value μ_i
  - $D_1, \ldots, D_k$ are independent standardized sub-gaussian

- Regret of policy π is the (worst-case) difference between
  - $T\mu^*$, the expected reward of a clairvoyant who knows the true distributions
  - $\mathbb{E}[\sum \mu_{\pi_t}]$, the expected reward of policy π

$$R^\pi(T) = \sup_{D_1,\ldots,D_k} \{T\mu^* - \mathbb{E}[\sum \mu_{\pi_t}]\}$$

- Objective: minimize regret

$$R^*(T) = \inf_\pi R^\pi(T)$$

Literature on MAB

- Applications in Revenue Management and Pricing
  - **Dynamic pricing with unknown demand**: Besbes and Zeevi (2009), Besbes and Zeevi (2012), Broder and Rusmevichientong (2012), Harrison et al. (2012), Keskin and Zeevi (2014), Wang et al. (2014), Chen et al. (2015), den Boer (2015), Cohen et al. (2016), Ferreira et al. (2018), ...
  - **Dynamic assortment optimization with unknown demand**: Rusmevichientong et al. (2010), Saure and Zeevi (2013), Agrawal et al. (2017a, b), Cheung and Simchi-Levi (2017), Chen et al. (2018), ...

- Stochastic MAB
  - **Algorithms**: successive elimination, UCB, Thompson sampling, ...
  - **Regret bounds**: $\Theta\left(\frac{\log T}{\Delta}\right)$ distribution-dependent bound (Lai and Robbins 1985), $\widetilde{\Theta}(\sqrt{T})$ distribution-free bound (Auer et al. 2002)

Literature on MAB

- Applications in Revenue Management and Pricing
  - **Dynamic pricing with unknown demand**: Besbes and Zeevi (2009), Besbes and Zeevi (2012), Broder and Rusmevichientong (2012), Harrison et al. (2012), Keskin and Zeevi (2014), Wang et al. (2014), Chen et al. (2015), den Boer (2015), Cohen et al. (2016), Ferreira et al. (2018), ...
  - **Dynamic assortment optimization with unknown demand**: Rusmevichientong et al. (2010), Saure and Zeevi (2013), Agrawal et al. (2017a, b), Cheung and Simchi-Levi (2017), Chen et al. (2018), ...

- Stochastic MAB
  - **Algorithms**: successive elimination, UCB, Thompson sampling, ...
  - **Regret bounds**: $\Theta\left(\frac{\log T}{\Delta}\right)$ distribution-dependent bound (Lai and Robbins 1985), $\tilde{\Theta}(\sqrt{T})$ distribution-free bound (Auer et al. 2002)

Motivation

- ▶ MAB deals with the "exploration-exploitation" trade-off
- ▶ Both exploration (i.e., acquiring new information) and exploitation (i.e., optimizing decisions based on up-to-date information) require switching
- ▶ In reality, switching between different arms is usually costly, and policies with limited switches are preferred
- ▶ Decision makers usually face strict limits on switching
  - E.g., Cheung et al. (2017), collaboration with **Groupon**

Motivating Example – Dynamic Pricing

- Single product, unlimited inventory, T rounds of sales
- Candidate prices: $p_1, \ldots, p_k$
- Price changes lead to negative customer feedback, and may induce undesirable strategic customer behavior
- Changing price from p_i to p_j incurs a cost $c_{i,j}$ $(i \neq j)$
- The seller wants to limit the total cost of price changes within a *switching budget S*
- For example:
  - $c_{i,j} = 1$, then S limits the number of price changes
  - $c_{i,j} = \mathbb{I}\{|p_i - p_j| > \eta\}$, then S limits the number of large price changes
  - $c_{i,j} = |p_i - p_j|$, then S limits the total price shift

Bandits with Switching Constraints (BwSC)

- Problem formulation
  - MAB subject to a Switching Constraint

- Switching constraint:
  - A switching budget, S, is the maximum amount of the total switching cost
  - Once the total switching cost exceeds the switching budget S, the decision-maker cannot switch her actions any more

Let $R_S^*(T)$ denote the optimal regret in BwSC

Talk Outline

- **Motivation**
  - Practical & Theoretical Motivation
- **MAB with Unit Switching Cost**
  - Limited Number of Switches
- **Extensions**
  - General Switching Costs
  - Inventory Constraints

S-Switch Successive Elimination (SS-SE) Policy

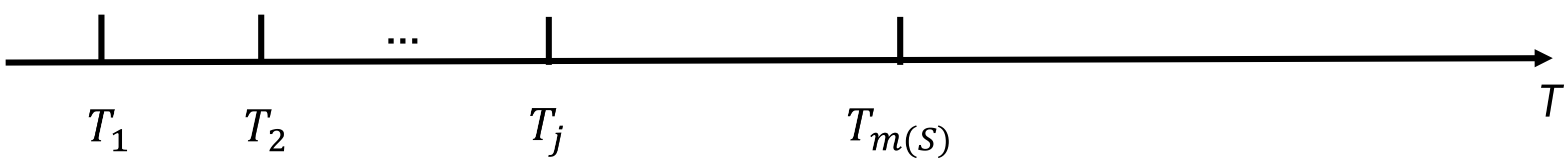
- Initiation
- Iterations
- Switching rules

The SS-SE Policy: Initialization

- Calculate an index

$$m(S) = \left\lfloor \frac{S-1}{k-1} \right\rfloor$$

- Partition the entire horizon into $(m(S)+1)$ intervals

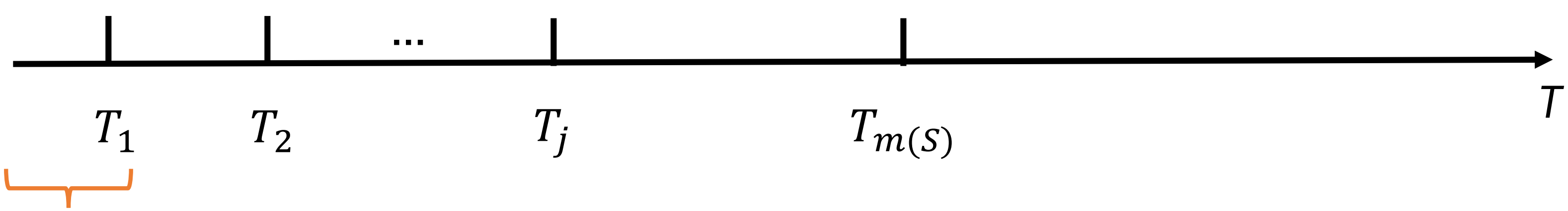


$$T_j \asymp T^{\frac{2-2^{1-j}}{2-2^{-m(S)}}}, \qquad j = 1,2, \dots m(S)+1$$

- Interval j is associated with an active set A_j $(j = 1, \dots, m(S))$
  - Let $A_1 = \{1, \dots, k\}$
  - A_{j+1} will be determined at the end of interval j

The SS-SE Policy: Iterations

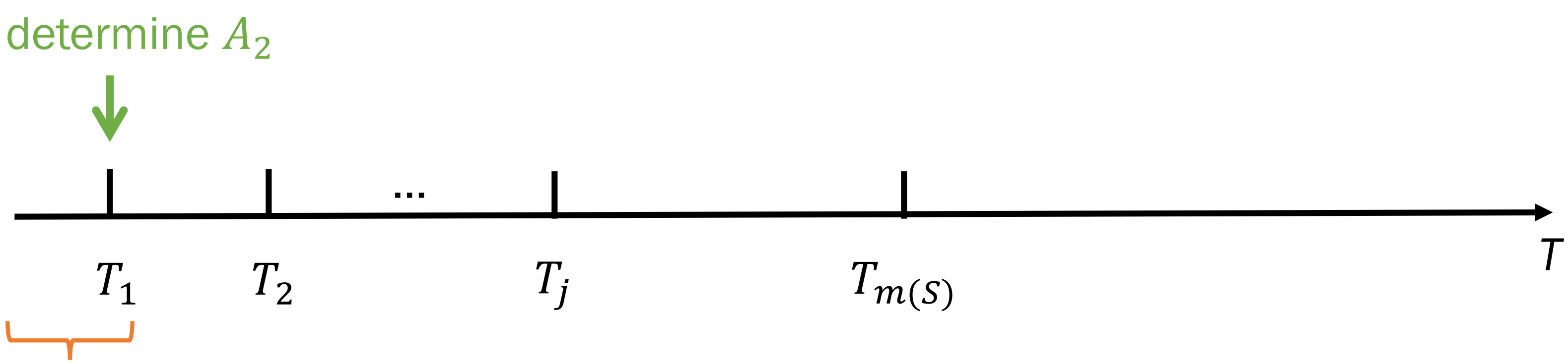
- Within each interval: pre-determined decisions



Within interval 1:
choose each action in A_1 for $\frac{T_1}{k}$ consecutive rounds

The SS-SE Policy: Iterations

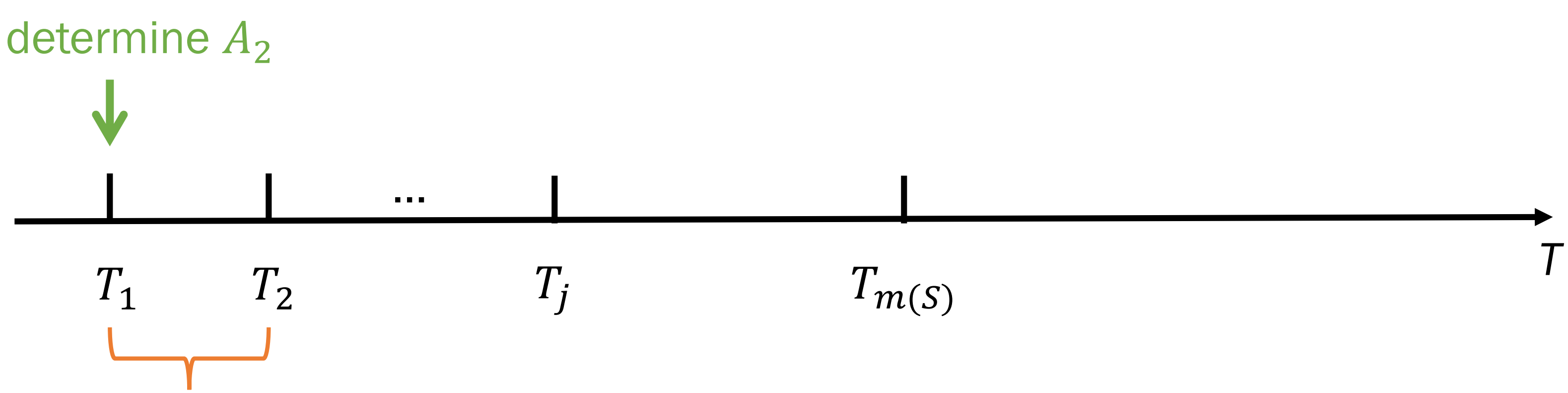
- Within each interval: pre-determined decisions
- At the end of each interval: eliminate ineffective actions based on confidence intervals



Within interval 1:
choose each action in A_1 for $\frac{T_1}{k}$ consecutive rounds

The SS-SE Policy: Iterations

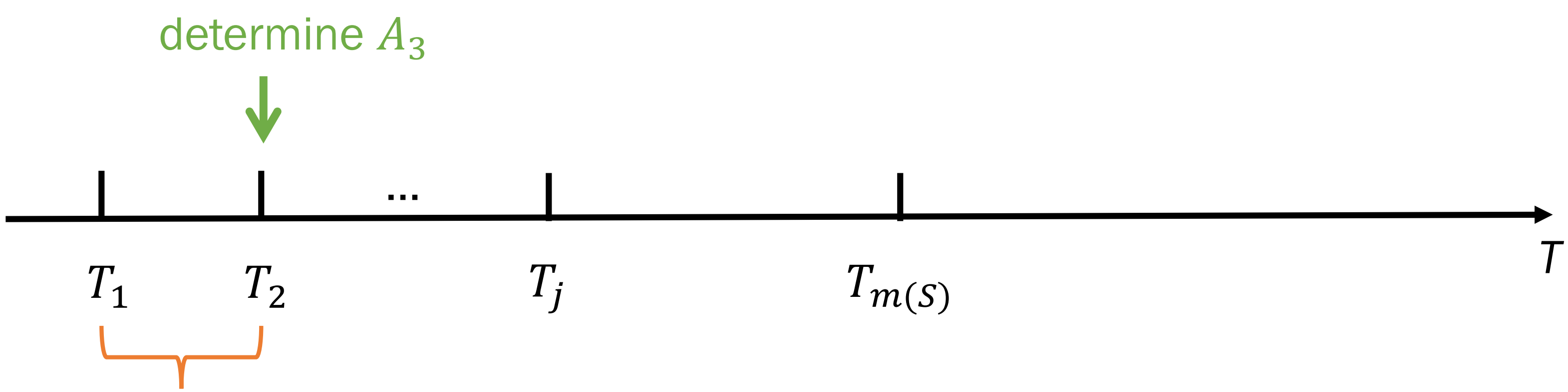
- Within each interval: pre-determined decisions
- At the end of each interval: eliminate ineffective actions based on confidence intervals



Within interval 2:
choose each action in A_2 for $\frac{T_2 - T_1}{|A_2|}$ consecutive rounds

The SS-SE Policy: Iterations

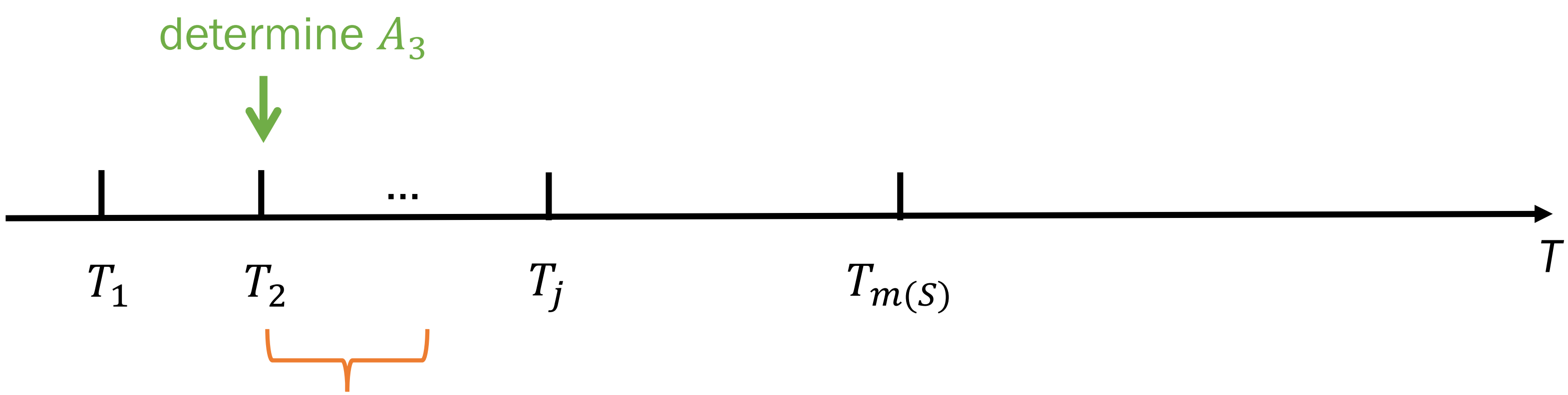
- Within each interval: pre-determined decisions
- At the end of each interval: eliminate ineffective actions based on confidence intervals



Within interval 2:
choose each action in A_2 for $\frac{T_2 - T_1}{|A_2|}$ consecutive rounds

The SS-SE Policy: Iterations

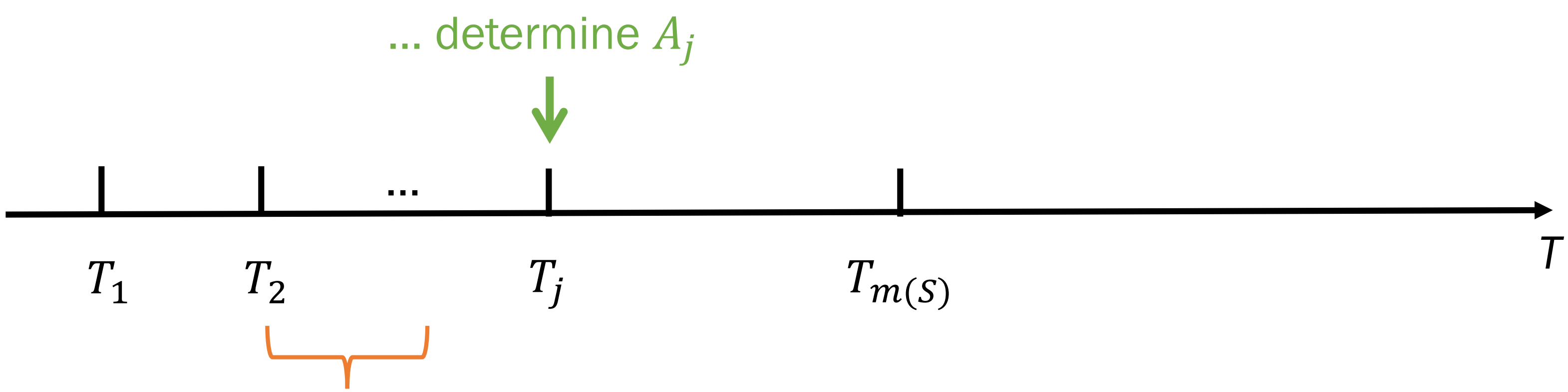
- Within each interval: pre-determined decisions
- At the end of each interval: eliminate ineffective actions based on confidence intervals



Within interval 3:
choose each action in A_3 for $\frac{T_3 - T_2}{|A_3|}$ consecutive rounds

The SS-SE Policy: Iterations

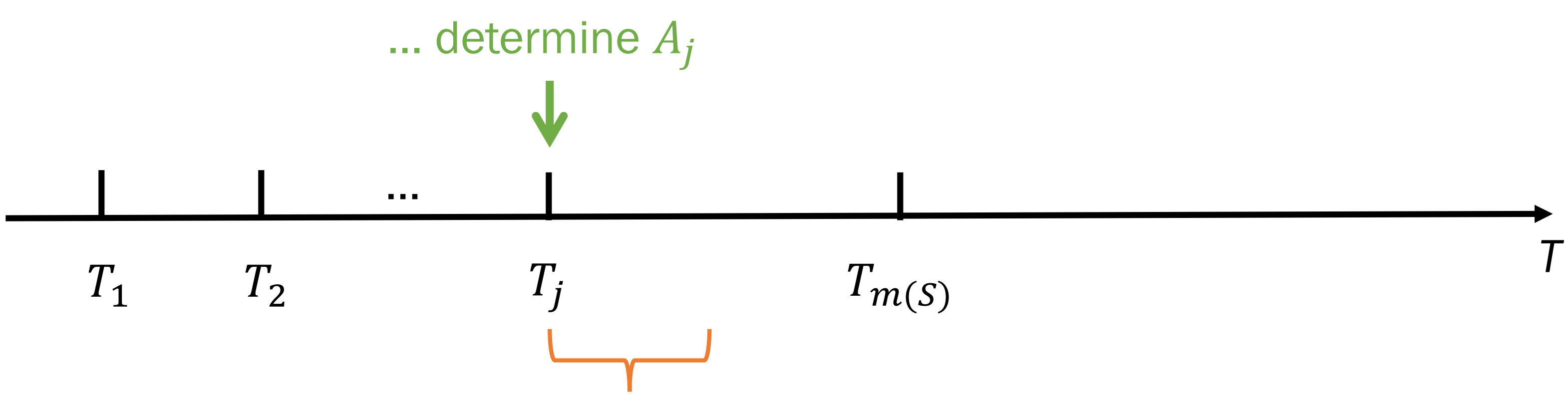
- Within each interval: pre-determined decisions
- At the end of each interval: eliminate ineffective actions based on confidence intervals



Within interval 3:
choose each action in A_3 for $\frac{T_3 - T_2}{|A_3|}$ consecutive rounds

The SS-SE Policy: Iterations

- Within each interval: pre-determined decisions
- At the end of each interval: eliminate ineffective actions based on confidence intervals

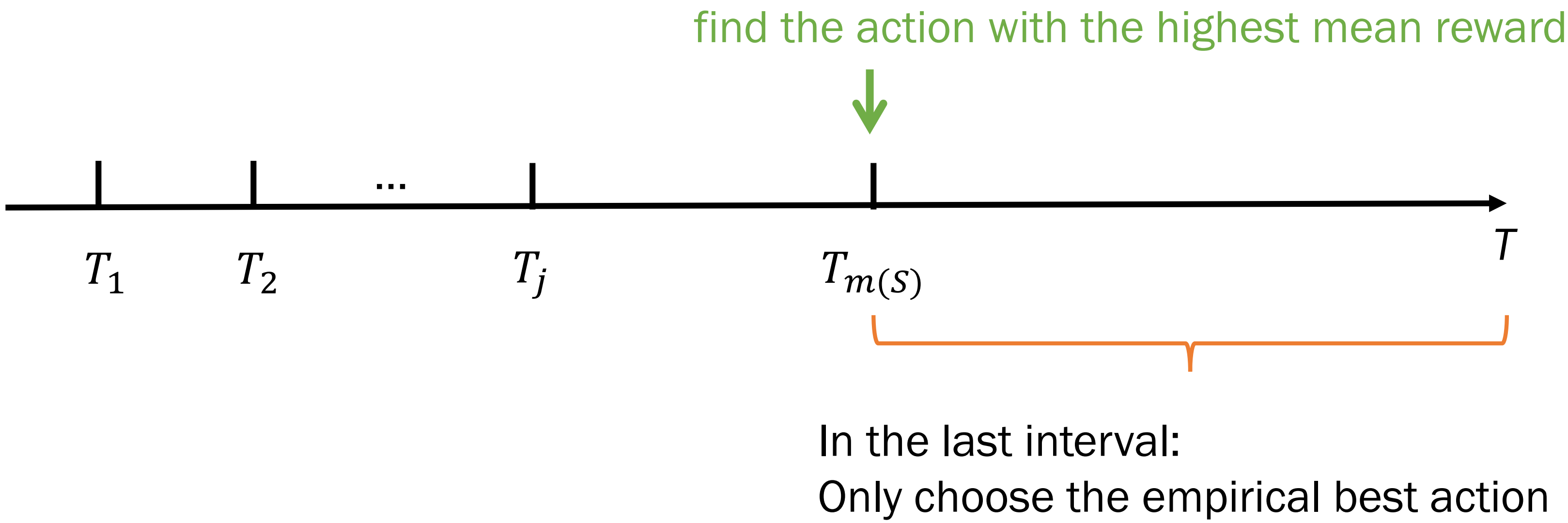


Within interval j:

choose each action in A_j for $\frac{T_j - T_{j-1}}{|A_j|}$ consecutive rounds

The SS-SE Policy: Iterations

- In the last interval: only choose the empirical best action



The SS-SE Policy: Switching Rule

At the start of each interval l:

- If the last action chosen in interval $l-1$ is still in A_l :
  - Interval l starts from this action
    - No switch happens between interval $l-1$ and interval l

- If the last action chosen in interval $l-1$ is eliminated from A_l :
  - Interval l starts from an arbitrary action in A_l
    - One switch happens between interval $l-1$ and interval l

The total number of switches is at most

$$m(S)(k-1)+1 \leq S$$

Upper Bound on Regret

▶ Regret of the SS-SE policy:

$$\boldsymbol{R^{\pi}(T) = \tilde{O}\left(T^{\frac{1}{2-2^{-m(S)}}}\right)}$$

Where $m(S) = \left\lfloor \frac{S-1}{k-1} \right\rfloor$.

▶ Recall $R_S^*(T)$ is the optimal regret of the BwSC problem

$$\boldsymbol{R_S^*(T) = \tilde{O}\left(T^{\frac{1}{2-2^{-\lfloor (S-1)/(k-1) \rfloor}}}\right)}$$

▶ UB decreases doubly exponentially with $m(S)$.

▶ $O(\log\log T)$ switches are enough for achieving $\tilde{O}(\sqrt{T})$ regret

Can We Do Better?

- Example: 11 actions, 20 switching budget, 100,000 rounds

  Since $m(S) = \left\lfloor \frac{20-1}{11-1} \right\rfloor$=1

  - Only 11 switches occur
  - Only inquire data once

- Two issues:
  - Unused switching budget
  - Low adaptivity

Lower Bound on Regret

- For any k, S, for any S-switching-budget policy π,

$$\boldsymbol{R^{\pi}(T) = \tilde{\Omega}\left(T^{\frac{1}{2-2^{-m(S)}}}\right)}$$

Where $m(S) = \left\lfloor \frac{S-1}{k-1} \right\rfloor$.

- This is a much stronger lower bound than finding a counter-example for specific k and S

Lower Bound on Regret

- For any k, S, for any S-switching-budget policy π,

$$\boldsymbol{R^{\pi}(T) = \widetilde{\Omega}\left(T^{\frac{1}{2-2^{-m(S)}}}\right)}$$

Where $m(S) = \left\lfloor \frac{S-1}{k-1} \right\rfloor$.

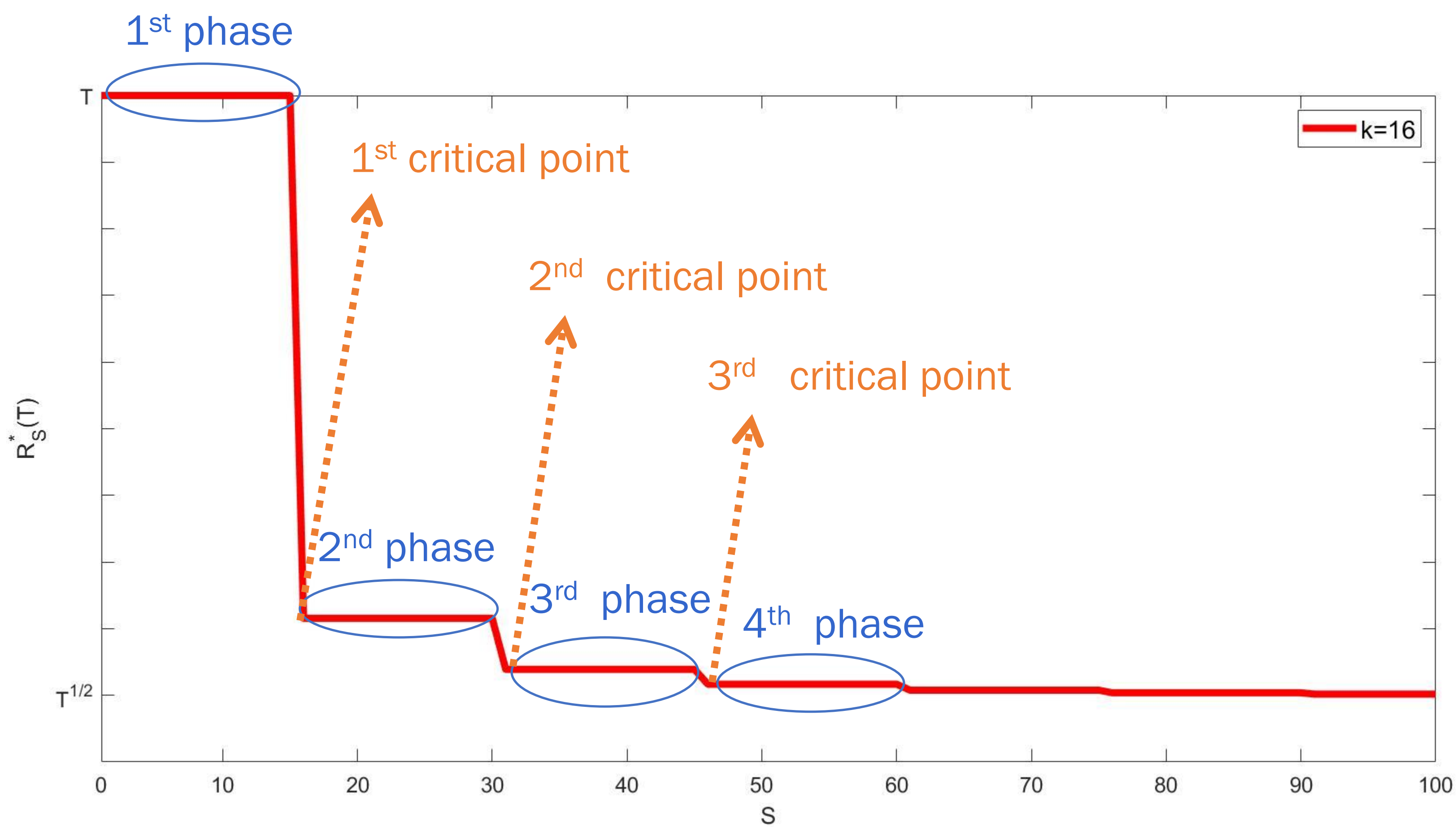
- Corollary

$$R_S^*(T) = \widetilde{\Theta}\left(T^{\frac{1}{2-2^{-\lfloor (S-1)/(k-1) \rfloor}}}\right)$$

Leads to some surprising behavior, namely, phase transitions and cyclic phenomena

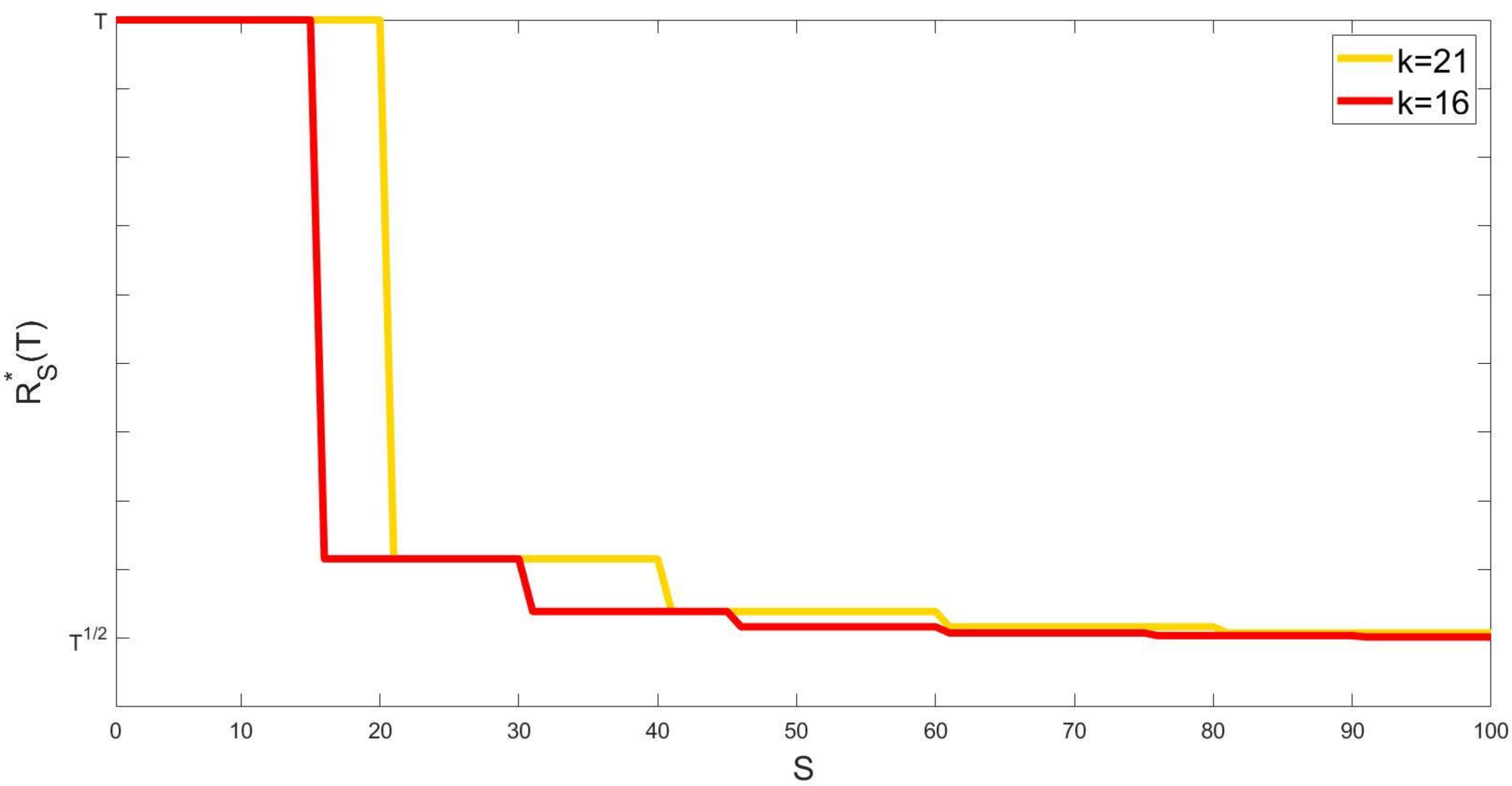
Phase Transitions

- Regret as a function of switching budget, S



Phase Transitions

- Regret as a function of switching budget, S



Phase Transitions

Table 1 Regret as a Function of Switching Budget

S	$[0, k)$	$[k, 2k-1)$	$[2k-1, 3k-2)$	$[3k-2, 4k-3)$	$[4k-3, 5k-4)$	$[5k-4, 6k-5)$
$R_S^*(T)$	$\tilde{\Theta}(T)$	$\tilde{\Theta}(T^{2/3})$	$\tilde{\Theta}(T^{4/7})$	$\tilde{\Theta}(T^{8/15})$	$\tilde{\Theta}(T^{16/31})$	$\tilde{\Theta}(T^{32/63})$
$R_S^*(T)/R_\infty^*(T)$	$\tilde{\Theta}(T^{1/2})$	$\tilde{\Theta}(T^{1/6})$	$\tilde{\Theta}(T^{1/14})$	$\tilde{\Theta}(T^{1/30})$	$\tilde{\Theta}(T^{1/62})$	$\tilde{\Theta}(T^{1/128})$

Consider the case of $k = 16$

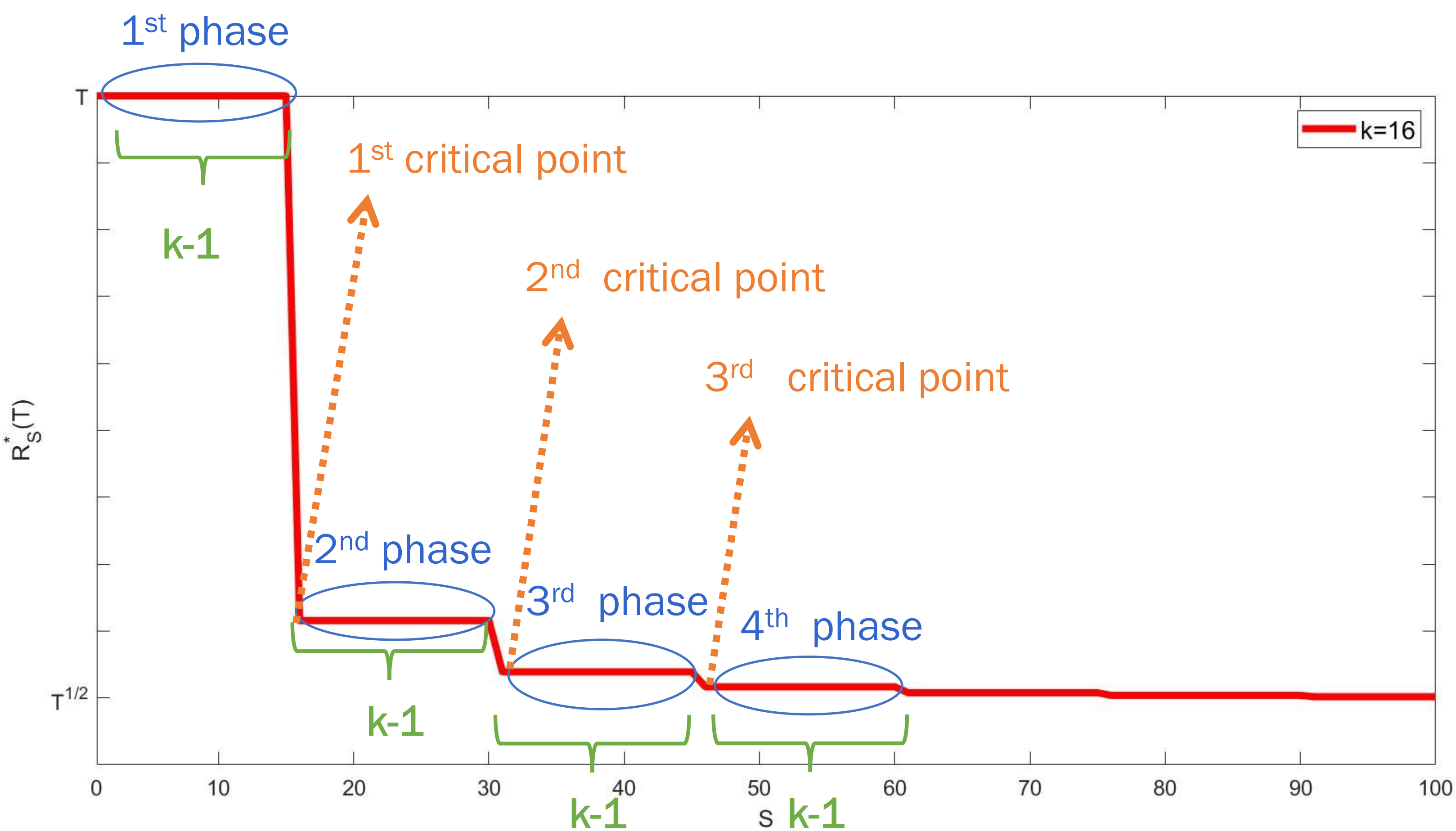
- When $S = 1,2, \dots, 15$
  - Regret is $\tilde{\Theta}(T)$
- When $S = 16,17, \dots, 30$
  - Regret is $\tilde{\Theta}(T^{2/3})$
- When $S = 31,32, \dots, 45$
  - Regret is $\tilde{\Theta}(T^{4/7})$
- When $S = 46,47, \dots, 60$
  - Regret is $\tilde{\Theta}(T^{8/15})$

More switching budget does not mean lower regret

Cyclic Phenomena

- The length of each phase

budget cycle



Cyclic Phenomena

Table 1 Regret as a Function of Switching Budget

S	$[0,k)$	$[k,2k-1)$	$[2k-1,3k-2)$	$[3k-2,4k-3)$	$[4k-3,5k-4)$	$[5k-4,6k-5)$
$R_S^*(T)$	$\tilde{\Theta}(T)$	$\tilde{\Theta}(T^{2/3})$	$\tilde{\Theta}(T^{4/7})$	$\tilde{\Theta}(T^{8/15})$	$\tilde{\Theta}(T^{16/31})$	$\tilde{\Theta}(T^{32/63})$
$R_S^*(T)/R_\infty^*(T)$	$\tilde{\Theta}(T^{1/2})$	$\tilde{\Theta}(T^{1/6})$	$\tilde{\Theta}(T^{1/14})$	$\tilde{\Theta}(T^{1/30})$	$\tilde{\Theta}(T^{1/62})$	$\tilde{\Theta}(T^{1/128})$

- The length of each phase is always $k-1$, independent of S and T
- 3 or 4 budget cycles are enough to achieve close-to-optimal regret
- Cyclic Phenomena seem counter-intuitive, since as one moves from phase to phase, one can
  - conduct more statistical tests;
  - eliminate more ineffective actions;
  - reduce the length of each phase.

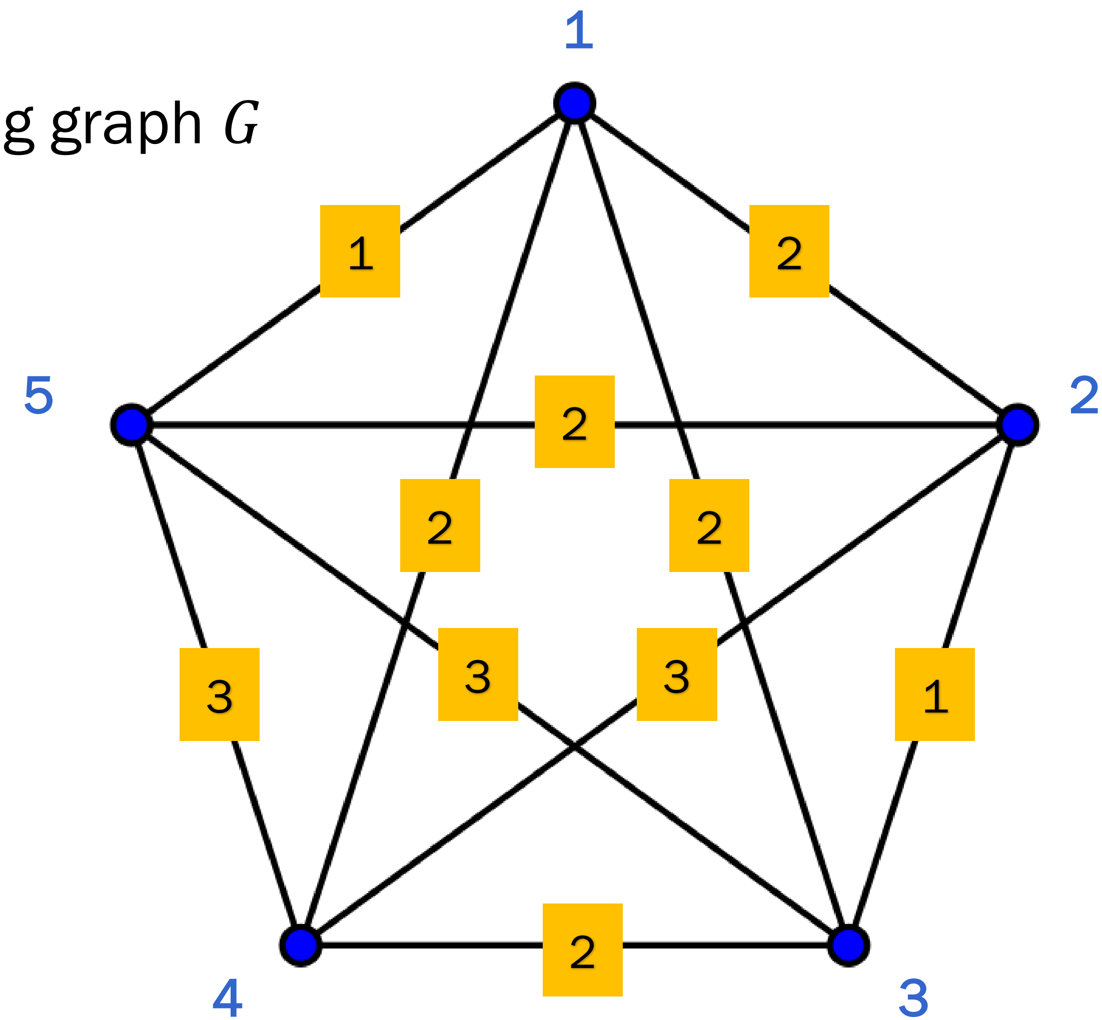
  On the surface, this suggests that the budget cycle may be a quantity decreasing with S

Talk Outline

- **Motivation**
  - **Practical & Theoretical Motivation**
- **MAB with Unit Switching Cost**
  - **Limited Number of Switches**
- **Extension**
  - **General Switching Costs**

Graph Representation

Switching graph G



Regret Bounds

- Let H denote the length of the shortest Hamiltonian path
- Upper bound on regret of the HS-SE policy

$$\boldsymbol{R^{\pi}(T) = \widetilde{O}\left(T^{\frac{1}{2-2^{-m_G^U(S)}}}\right)}$$

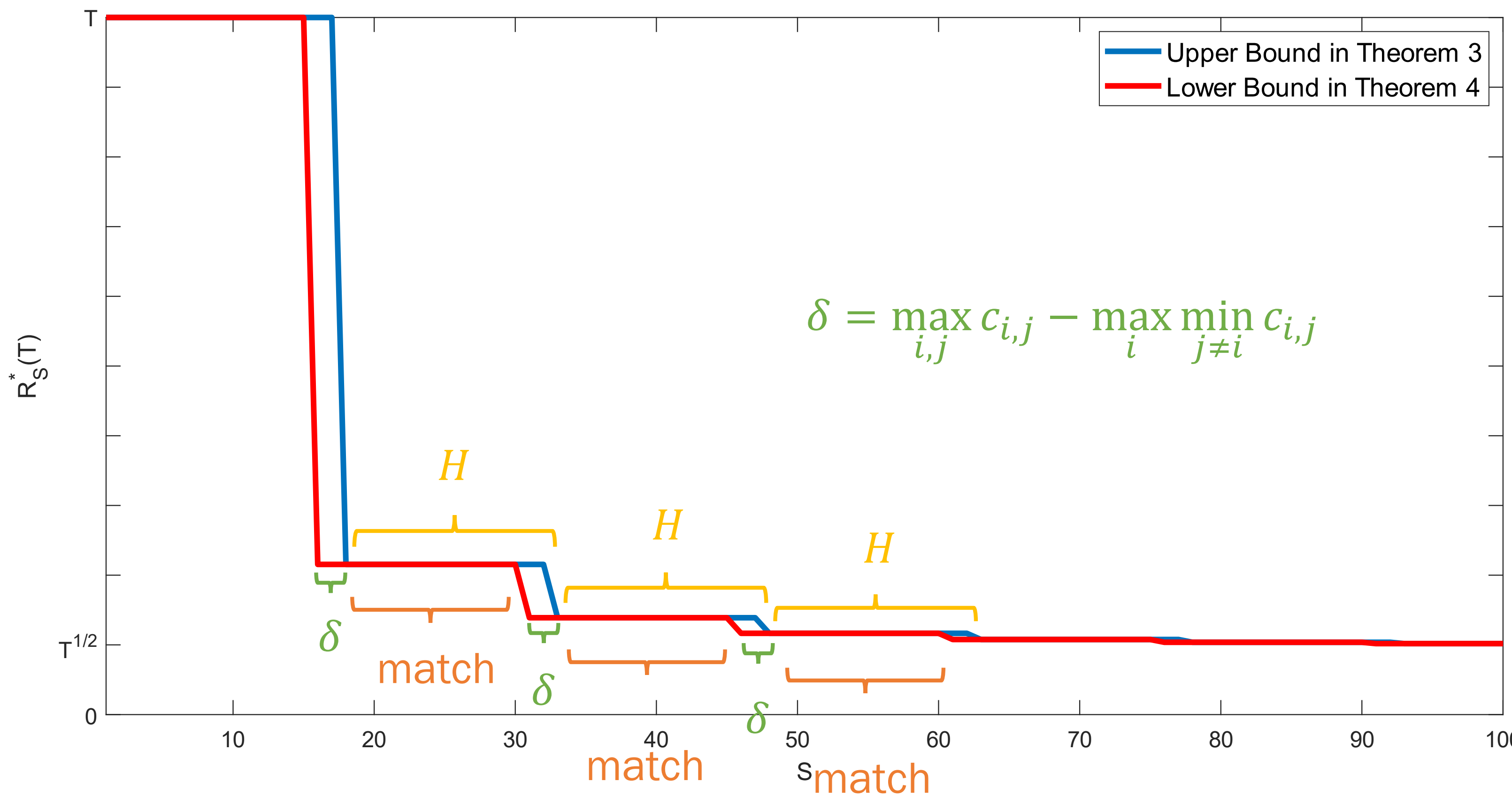
Where $m_G^U(S) = \left\lfloor \frac{S - \max_{i,j} c_{i,j}}{H} \right\rfloor$

- Lower bound on regret of any S-switching-budget policy

$$\boldsymbol{R^{\pi}(T) = \widetilde{\Omega}\left(T^{\frac{1}{2-2^{-m_G^L(S)}}}\right)}$$

Where $m_G^L(S) = \left\lfloor \frac{S - \max_{i} \min_{j \neq i} c_{i,j}}{H} \right\rfloor$

The Bounds are Close to Each Other



Summary

- MAB with Limited Number of Switches exhibits phase transitions and cyclic phenomena

- MAB with General Switching Cost maintains phase transitions but not cyclic phenomena

Thank you!