Scaling Multi-Armed Bandit Algorithms

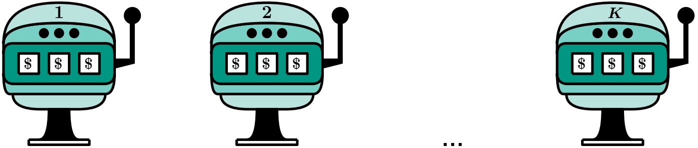
25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD ’19)

Edouard Fouché*, Junpei Komiyama** & Klemens Böhm* | August 8, 2019

* KARLSRUHE INSTITUTE OF TECHNOLOGY (KIT), ** THE UNIVERSITY OF TOKYO

This talk is about the Multi-Armed Bandit (MAB)

Karlsruhe Institute of Technology

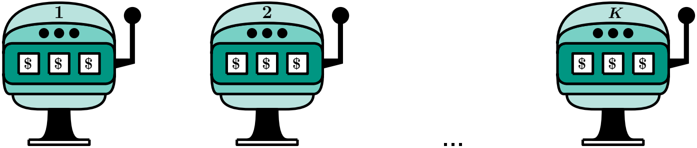


The MAB is a well-known model for sequential decision making.

- Work on bandits traces to back to (Thompson, 1933)
- Theoretical guarantees remained unknown until recently
  - (Auer et al., 1995, 2002; Garivier and Moulines, 2008; Kaufmann et al., 2012)

We present an extension: The **Scaling** Multi-Armed Bandit (S-MAB)

This talk is about the Multi-Armed Bandit (MAB)

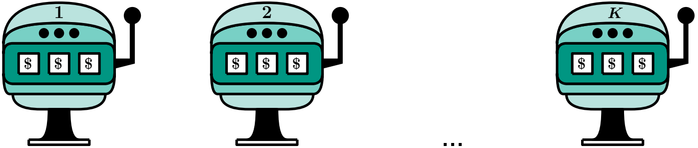


The MAB is a well-known model for sequential decision making.

- Work on bandits traces to back to (Thompson, 1933)
- Theoretical guarantees remained unknown until recently
  - (Auer et al., 1995, 2002; Garivier and Moulines, 2008; Kaufmann et al., 2012)

We present an extension: The **Scaling** Multi-Armed Bandit (S-MAB)

This talk is about the Multi-Armed Bandit (MAB)

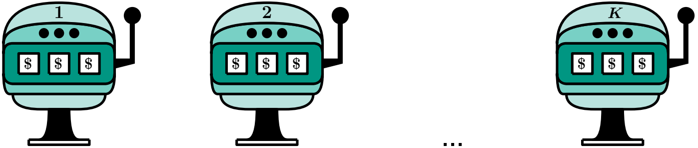


The MAB is a well-known model for sequential decision making.

- Work on bandits traces to back to (Thompson, 1933)
- Theoretical guarantees remained unknown until recently
  - (Auer et al., 1995, 2002; Garivier and Moulines, 2008; Kaufmann et al., 2012)

We present an extension: The **Scaling** Multi-Armed Bandit (S-MAB)

This talk is about the Multi-Armed Bandit (MAB)



The MAB is a well-known model for sequential decision making.

- Work on bandits traces to back to (Thompson, 1933)
- Theoretical guarantees remained unknown until recently
  - (Auer et al., 1995, 2002; Garivier and Moulines, 2008; Kaufmann et al., 2012)

We present an extension: The **Scaling** Multi-Armed Bandit (S-MAB)

Use Case: Data Stream Monitoring

Bioliq power plant at KIT[1]

- **Bio**mass-to-**Liq**uids
- Pyrolysis: Biomass $\rightarrow$ Biogas

A high-dimensional data stream:

- $>$ 800 sensors
- 1 new data point per second

The Bioliq® power plant

Our goal: Monitor highly-correlated pairs in this stream

Code & data: https://github.com/edouardfouche/S-MAB

[1] https://bioliq.de

Use Case: Data Stream Monitoring

Bioliq power plant at KIT[1]

- **Bio**mass-to-**Liq**uids
- Pyrolysis: Biomass $\rightarrow$ Biogas

A high-dimensional data stream:

- $>$ 800 sensors
- 1 new data point per second

The Bioliq® power plant

Our goal: Monitor highly-correlated pairs in this stream

Code & data: https://github.com/edouardfouche/S-MAB

[1] https://bioliq.de

Use Case: Data Stream Monitoring

Bioliq power plant at KIT[1]

- **Bio**mass-to-**Liq**uids
- Pyrolysis: Biomass $\rightarrow$ Biogas

A high-dimensional data stream:

- $>$ 800 sensors
- 1 new data point per second

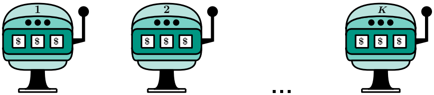
The Bioliq® power plant

Our goal: Monitor highly-correlated pairs in this stream

Code & data: https://github.com/edouardfouche/S-MAB

[1] https://bioliq.de

The "Classical" MAB



Let there be a set of K arms, $[K] = \{1, \ldots, K\}$.

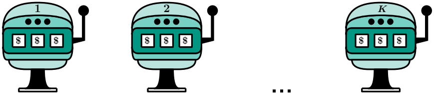
- Each $i \in [K]$ is associated to a Bernouilli distribution $\mathcal{B}(\mu_i)$; μ_i unknown.

At each round $t = 1, \ldots, T$:

- The forecaster chooses **one** arm $i \in [K]$
- Then, she observes a reward $X_t \sim \mathcal{B}(\mu_i)$
- She updates her estimation $\hat{\mu}_i$ of μ_i

The goal of the forecaster is to maximize her total reward, i.e., $\sum_{t=1}^{T} X_t$

The “Classical” MAB



Let there be a set of K arms, $[K] = \{1, \dots, K\}$.

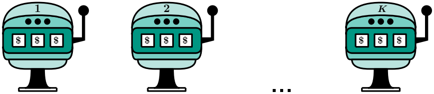
- Each $i \in [K]$ is associated to a Bernouilli distribution $\mathcal{B}(\mu_i)$; μ_i unknown.

At each round $t = 1, \dots, T$:

- The forecaster chooses **one** arm $i \in [K]$
- Then, she observes a reward $X_t \sim \mathcal{B}(\mu_i)$
- She updates her estimation $\hat{\mu}_i$ of μ_i

The goal of the forecaster is to maximize her total reward, i.e., $\sum_{t=1}^{T} X_t$

The “Classical” MAB



Let there be a set of K arms, $[K] = \{1, \ldots, K\}$.

- Each $i \in [K]$ is associated to a Bernouilli distribution $\mathcal{B}(\mu_i)$; μ_i unknown.

At each round $t = 1, \ldots, T$:

- The forecaster chooses **one** arm $i \in [K]$
- Then, she observes a reward $X_t \sim \mathcal{B}(\mu_i)$
- She updates her estimation $\hat{\mu}_i$ of μ_i

The goal of the forecaster is to maximize her total reward, i.e., $\sum_{t=1}^{T} X_t$

The MAB with Multiple Plays (MP-MAB)

MP-MAB: A model for online subset selection.

The forecaster plays $L > 1$ arms per round.

- Extension discussed in (Uchiya et al., 2010; Komiyama et al., 2015).

Problem: L is fixed as an external parameter.

- Typically, playing an arm is associated to a cost
- User needs to set L
  - L is too large $\rightarrow$ Cost $>$ Reward
  - L is too small $\rightarrow$ loss of potential gain

An "efficient" number of plays is unknown a priori!

Non-Static: The distribution parameters $\mu_1, \ldots, \mu_K$ may vary over time.

The MAB with Multiple Plays (MP-MAB)

MP-MAB: A model for online subset selection.

The forecaster plays $L > 1$ arms per round.

- Extension discussed in (Uchiya et al., 2010; Komiyama et al., 2015).

Problem: L is fixed as an external parameter.

- Typically, playing an arm is associated to a cost
- User needs to set L
  - L is too large $\rightarrow$ Cost $>$ Reward
  - L is too small $\rightarrow$ loss of potential gain

An “efficient” number of plays is unknown a priori!

Non-Static: The distribution parameters $\mu_1, \ldots, \mu_K$ may vary over time.

The MAB with Multiple Plays (MP-MAB)

MP-MAB: A model for online subset selection.

The forecaster plays $L > 1$ arms per round.

- Extension discussed in (Uchiya et al., 2010; Komiyama et al., 2015).

Problem: L is fixed as an external parameter.

- Typically, playing an arm is associated to a cost
- User needs to set L
  - L is too large $\rightarrow$ Cost $>$ Reward
  - L is too small $\rightarrow$ loss of potential gain

An "efficient" number of plays is unknown a priori!

Non-Static: The distribution parameters $\mu_1, \ldots, \mu_K$ may vary over time.

The MAB with Multiple Plays (MP-MAB)

MP-MAB: A model for online subset selection.

The forecaster plays $L > 1$ arms per round.

- Extension discussed in (Uchiya et al., 2010; Komiyama et al., 2015).

Problem: L is fixed as an external parameter.

- Typically, playing an arm is associated to a cost
- User needs to set L
  - L is too large $\rightarrow$ Cost $>$ Reward
  - L is too small $\rightarrow$ loss of potential gain

An "efficient" number of plays is unknown a priori!

Non-Static: The distribution parameters $\mu_1, \ldots, \mu_K$ may vary over time.

The MAB with Multiple Plays (MP-MAB)

MP-MAB: A model for online subset selection.

The forecaster plays $L > 1$ arms per round.

- Extension discussed in (Uchiya et al., 2010; Komiyama et al., 2015).

Problem: L is fixed as an external parameter.

- Typically, playing an arm is associated to a cost
- User needs to set L
  - L is too large $\rightarrow$ Cost $>$ Reward
  - L is too small $\rightarrow$ loss of potential gain

An “efficient” number of plays is unknown a priori!

Non-Static: The distribution parameters $\mu_1, \ldots, \mu_K$ may vary over time.

The MAB with Multiple Plays (MP-MAB)

MP-MAB: A model for online subset selection.

The forecaster plays $L > 1$ arms per round.

- Extension discussed in (Uchiya et al., 2010; Komiyama et al., 2015).

Problem: L is fixed as an external parameter.

- Typically, playing an arm is associated to a cost
- User needs to set L
  - L is too large $\rightarrow$ Cost $>$ Reward
  - L is too small $\rightarrow$ loss of potential gain

An "efficient" number of plays is unknown a priori!

Non-Static: The distribution parameters $\mu_1, \ldots, \mu_K$ may vary over time.

Real-world Applications

Data Stream Monitoring

- Too much monitoring is a waste of resources
- But events of interest might go unnoticed
- **arm**: statistics, **round**: timestep, **reward**: interest

But also:

- **Online Advertisement**
- **Financial Investment**

Real-world Applications

Data Stream Monitoring

- Too much monitoring is a waste of resources
- But events of interest might go unnoticed
- **arm**: statistics, **round**: timestep, **reward**: interest

But also:

- **Online Advertisement**
- **Financial Investment**

Real-world Applications

Karlsruhe Institute of Technology

Data Stream Monitoring

- Too much monitoring is a waste of resources
- But events of interest might go unnoticed
- **arm**: statistics, **round**: timestep, **reward**: interest

But also:

- **Online Advertisement**
- **Financial Investment**

Problem Definition

Multiple-play MAB with efficiency constraint

- Let be $I_t \subset [K]$ the set of arms played at time t, with $|I_t| = L_t$
- $S_i(t)$ is the sum of the rewards from arm i up to time t

Goal: Maximize the reward subject to cost constraint

$$\max_{I_t \subset [K]} \sum_{i \in I_t} S_i(t) \quad s.t. \quad \eta_t = \frac{\sum_{i \in I_t} \mu_i}{L_t} > \eta^* \tag{1}$$

If the forecaster always chooses the top-L_t arms, then the problem is equivalent to finding the optimal number of plays L^*:

$$L^* = \max_{1 \leq L \leq K} L \quad s.t. \quad \frac{\sum_{i=1}^{L} \mu_i}{L} > \eta^* \tag{2}$$

Problem Definition

Multiple-play MAB with efficiency constraint

- Let be $I_t \subset [K]$ the set of arms played at time t, with $|I_t| = L_t$
- $S_i(t)$ is the sum of the rewards from arm i up to time t

Goal: Maximize the reward subject to cost constraint

$$\max_{I_t \subset [K]} \sum_{i \in I_t} S_i(t) \quad s.t. \quad \eta_t = \frac{\sum_{i \in I_t} \mu_i}{L_t} > \eta^* \tag{1}$$

If the forecaster always chooses the top-L_t arms, then the problem is equivalent to finding the optimal number of plays L^*:

$$L^* = \max_{1 \leq L \leq K} L \quad s.t. \quad \frac{\sum_{i=1}^{L} \mu_i}{L} > \eta^* \tag{2}$$

General Scaling Multi-Armed Bandit

Two components: finding the top-L_t + finding L_t (new)

At each round $t = 1, \ldots, T$:

- 1. The forecaster chooses I_t with $|I_t| = L_t$, and observes a reward vector X_t
- 2. She updates her estimation $\hat{\mu}_i$ for $i \in I_t$
- 3. **She chooses L_{t+1} ($\rightarrow$ Scaling)**

There exists many approaches for steps $1, 2$:

- Thompson Sampling (TS) (Thompson, 1933; Kaufmann et al., 2012)
- UCB-type (Auer et al., 2002; Chen et al., 2016; Garivier and Cappé, 2011)

For step 3 $\rightarrow$ We introduce a "scaling policy" (see next slide)

General Scaling Multi-Armed Bandit

Karlsruhe Institute of Technology

Two components: finding the top-L_t + finding L_t (new)

At each round $t = 1, \ldots, T$:

- 1. The forecaster chooses I_t with $|I_t| = L_t$, and observes a reward vector X_t
- 2. She updates her estimation $\hat{\mu}_i$ for $i \in I_t$
- 3. **She chooses L_{t+1} ($\rightarrow$ Scaling)**

There exists many approaches for steps $1, 2$:

- Thompson Sampling (TS) (Thompson, 1933; Kaufmann et al., 2012)
- UCB-type (Auer et al., 2002; Chen et al., 2016; Garivier and Cappé, 2011)

For step 3 $\rightarrow$ We introduce a "scaling policy" (see next slide)

General Scaling Multi-Armed Bandit

Karlsruhe Institute of Technology

Two components: finding the top-L_t + finding L_t (new)

At each round $t = 1, \ldots, T$:

- 1. The forecaster chooses I_t with $|I_t| = L_t$, and observes a reward vector X_t
- 2. She updates her estimation $\hat{\mu}_i$ for $i \in I_t$
- 3. **She chooses L_{t+1} ($\rightarrow$ Scaling)**

There exists many approaches for steps $1, 2$:

- Thompson Sampling (TS) (Thompson, 1933; Kaufmann et al., 2012)
- UCB-type (Auer et al., 2002; Chen et al., 2016; Garivier and Cappé, 2011)

For step 3 $\rightarrow$ We introduce a “scaling policy” (see next slide)

General Scaling Multi-Armed Bandit

Karlsruhe Institute of Technology

Two components: finding the top-L_t + finding L_t (new)

At each round $t = 1, \ldots, T$:

- 1. The forecaster chooses I_t with $|I_t| = L_t$, and observes a reward vector X_t
- 2. She updates her estimation $\hat{\mu}_i$ for $i \in I_t$
- 3. **She chooses L_{t+1} ($\rightarrow$ Scaling)**

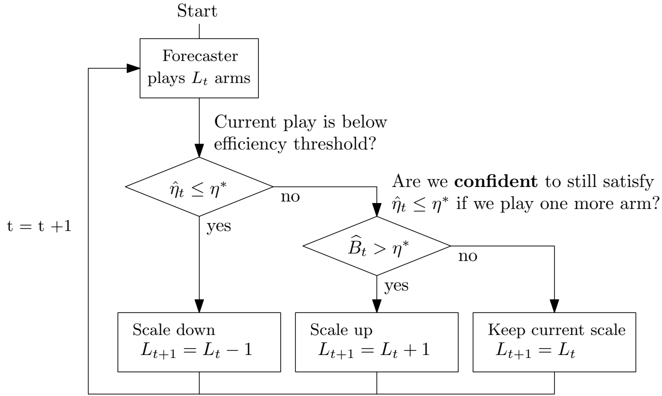
There exists many approaches for steps $1, 2$:

- Thompson Sampling (TS) (Thompson, 1933; Kaufmann et al., 2012)
- UCB-type (Auer et al., 2002; Chen et al., 2016; Garivier and Cappé, 2011)

For step 3 $\rightarrow$ We introduce a "scaling policy" (see next slide)

Scaling Policy

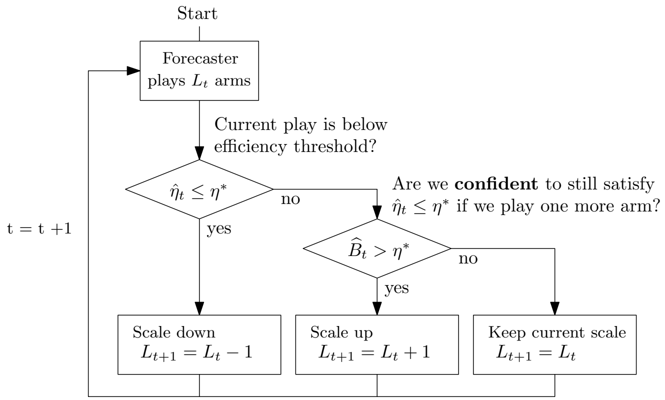
Kullback-Leibler Scaling (KL-S)



$\widehat{B}_t$ uses the Kullback-Leibler UCB (KL-UCB) index (Garivier and Cappé, 2011)

Scaling Policy

Kullback-Leibler Scaling (KL-S)



$\widehat{B}_t$ uses the Kullback-Leibler UCB (KL-UCB) index (Garivier and Cappé, 2011)

Regret

How to evaluate our approach?

We want to minimize the multiple-play regret:

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \left[\max_{\mathcal{I} \subseteq [K], |\mathcal{I}| = L_t} \sum_{i \in \mathcal{I}} \mu_i - \sum_{i \in I(t)} \mu_i \right]$$

also, we want to minimize the “pull” regret (new):

$$\mathrm{PReg}(T) = \sum_{t=1}^{T} |L^* - L_t|$$

$\mathrm{Reg}(T)$ is small $\rightarrow$ Top-L arms identification (step $1, 2$)
$\mathrm{PReg}(T)$ is small $\rightarrow$ Scaling converges to L^* (step 3)

Regret

How to evaluate our approach?

We want to minimize the multiple-play regret:

$$\text{Reg}(T) = \sum_{t=1}^{T} \left[\max_{\mathcal{I} \subseteq [K], |\mathcal{I}| = L_t} \sum_{i \in \mathcal{I}} \mu_i - \sum_{i \in I(t)} \mu_i \right]$$

also, we want to minimize the “pull” regret (new):

$$\text{PReg}(T) = \sum_{t=1}^{T} |L^* - L_t|$$

$\text{Reg}(T)$ is small $\rightarrow$ Top-L arms identification (step $1, 2$)
$\text{PReg}(T)$ is small $\rightarrow$ Scaling converges to L^* (step 3)

Regret

How to evaluate our approach?

We want to minimize the multiple-play regret:

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \left[\max_{\mathcal{I} \subseteq [K], |\mathcal{I}| = L_t} \sum_{i \in \mathcal{I}} \mu_i - \sum_{i \in I(t)} \mu_i \right]$$

also, we want to minimize the “pull” regret (new):

$$\mathrm{PReg}(T) = \sum_{t=1}^{T} |L^* - L_t|$$

$\mathrm{Reg}(T)$ is small $\rightarrow$ Top-L arms identification (step $1, 2$)
$\mathrm{PReg}(T)$ is small $\rightarrow$ Scaling converges to L^* (step 3)

Regret

How to evaluate our approach?

We want to minimize the multiple-play regret:

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \left[\max_{\mathcal{I} \subseteq [K], |\mathcal{I}| = L_t} \sum_{i \in \mathcal{I}} \mu_i - \sum_{i \in I(t)} \mu_i \right]$$

also, we want to minimize the “pull” regret (new):

$$\mathrm{PReg}(T) = \sum_{t=1}^{T} |L^* - L_t|$$

$\mathrm{Reg}(T)$ is small $\rightarrow$ Top-L arms identification (step $1, 2$)
$\mathrm{PReg}(T)$ is small $\rightarrow$ Scaling converges to L^* (step 3)

Main Result

Theorem (Logarithmic regret and pull regret)

*The Scaling MAB has **logarithmic regret** and **logarithmic pull regret**, provided that the underlying MAB has logarithmic regret, i.e., there exist two constants C_1, C_2 such that:*

$$\mathbb{E}[Reg(T)] \leq C_1 \log T \qquad \mathbb{E}[PReg(T)] \leq C_2 \log T \tag{3}$$

Scaling + "state-of-the-art MAB" $\rightarrow$ logarithmic regret/pull regret.

For example: Thompson Sampling (TS) and UCB-type bandits.

- Scaling + TS (Thompson, 1933) $\rightarrow$ S-TS
- Scaling + KL-UCB (Garivier and Cappé, 2011) $\rightarrow$ S-KL-UCB

Main Result

Theorem (Logarithmic regret and pull regret)

The Scaling MAB has ***logarithmic regret*** *and* ***logarithmic pull regret****, provided that the underlying MAB has logarithmic regret, i.e., there exist two constants C_1, C_2 such that:*

$$\mathbb{E}[Reg(T)] \leq C_1 \log T \qquad \mathbb{E}[PReg(T)] \leq C_2 \log T \tag{3}$$

Scaling + "state-of-the-art MAB" $\rightarrow$ logarithmic regret/pull regret.

For example: Thompson Sampling (TS) and UCB-type bandits.

- Scaling + TS (Thompson, 1933) $\rightarrow$ S-TS
- Scaling + KL-UCB (Garivier and Cappé, 2011) $\rightarrow$ S-KL-UCB

Main Result

Theorem (Logarithmic regret and pull regret)

The Scaling MAB has ***logarithmic regret*** *and* ***logarithmic pull regret****, provided that the underlying MAB has logarithmic regret, i.e., there exist two constants C_1, C_2 such that:*

$$\mathbb{E}[Reg(T)] \leq C_1 \log T \qquad \mathbb{E}[PReg(T)] \leq C_2 \log T \tag{3}$$

Scaling + “state-of-the-art MAB” $\rightarrow$ logarithmic regret/pull regret.

For example: Thompson Sampling (TS) and UCB-type bandits.

- Scaling + TS (Thompson, 1933) $\rightarrow$ S-TS
- Scaling + KL-UCB (Garivier and Cappé, 2011) $\rightarrow$ S-KL-UCB

Non-static Adaptation

Problem: The expectations μ_i of each arm $i \in [K]$ might change.

We use Adaptive Windowing (ADWIN) (Bifet and Gavaldà, 2007)

- Maintain $\hat{\mu}_i$ for each arm over a sliding window of adaptive length

$\rightarrow$ Scaling Thompson Sampling with ADWIN (S-TS-A)

Non-static Adaptation

Karlsruhe Institute of Technology

Problem: The expectations μ_i of each arm $i \in [K]$ might change.

We use Adaptive Windowing (ADWIN) (Bifet and Gavaldà, 2007)

- Maintain $\hat{\mu}_i$ for each arm over a sliding window of adaptive length

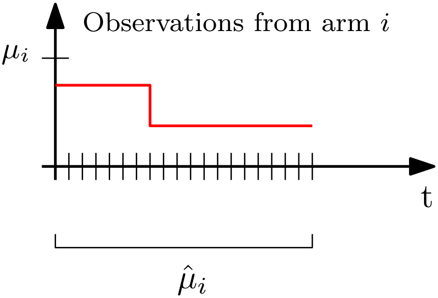
$\rightarrow$ Scaling Thompson Sampling with ADWIN (S-TS-A)

Non-static Adaptation

Problem: The expectations μ_i of each arm $i \in [K]$ might change.

We use Adaptive Windowing (ADWIN) (Bifet and Gavaldà, 2007)

- Maintain $\hat{\mu}_i$ for each arm over a sliding window of adaptive length



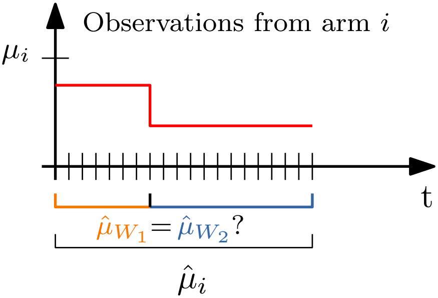
$\rightarrow$ Scaling Thompson Sampling with ADWIN (S-TS-A)

Non-static Adaptation

Problem: The expectations μ_i of each arm $i \in [K]$ might change.

We use Adaptive Windowing (ADWIN) (Bifet and Gavaldà, 2007)

- Maintain $\hat{\mu}_i$ for each arm over a sliding window of adaptive length



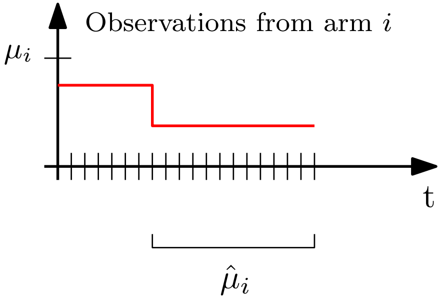
$\rightarrow$ Scaling Thompson Sampling with ADWIN (S-TS-A)

Non-static Adaptation

Problem: The expectations μ_i of each arm $i \in [K]$ might change.

We use Adaptive Windowing (ADWIN) (Bifet and Gavaldà, 2007)

- Maintain $\hat{\mu}_i$ for each arm over a sliding window of adaptive length



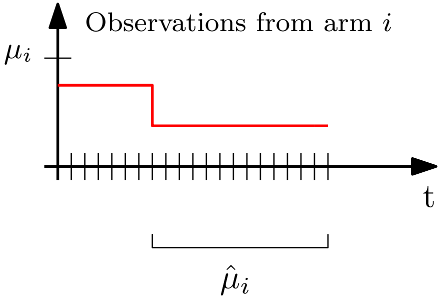
$\rightarrow$ Scaling Thompson Sampling with ADWIN (S-TS-A)

Non-static Adaptation

Problem: The expectations μ_i of each arm $i \in [K]$ might change.

We use Adaptive Windowing (ADWIN) (Bifet and Gavaldà, 2007)

- Maintain $\hat{\mu}_i$ for each arm over a sliding window of adaptive length



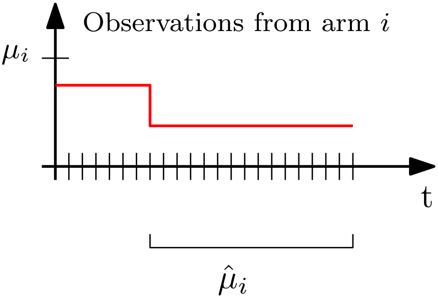
→ Scaling Thompson Sampling with ADWIN (S-TS-A)

Non-static Adaptation

Problem: The expectations μ_i of each arm $i \in [K]$ might change.

We use Adaptive Windowing (ADWIN) (Bifet and Gavaldà, 2007)

- Maintain $\hat{\mu}_i$ for each arm over a sliding window of adaptive length



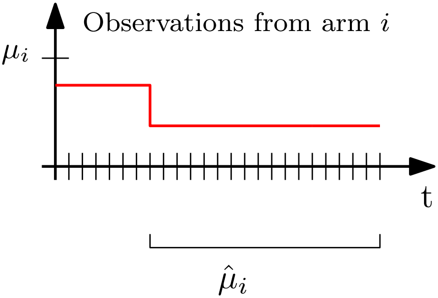
→ Scaling Thompson Sampling with ADWIN (S-TS-A)

Non-static Adaptation

Problem: The expectations μ_i of each arm $i \in [K]$ might change.

We use Adaptive Windowing (ADWIN) (Bifet and Gavaldà, 2007)

- Maintain $\hat{\mu}_i$ for each arm over a sliding window of adaptive length



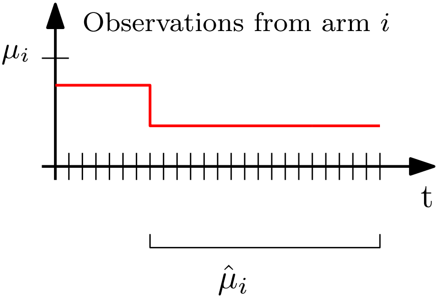
→ Scaling Thompson Sampling with ADWIN (S-TS-A)

Non-static Adaptation

Problem: The expectations μ_i of each arm $i \in [K]$ might change.

We use Adaptive Windowing (ADWIN) (Bifet and Gavaldà, 2007)

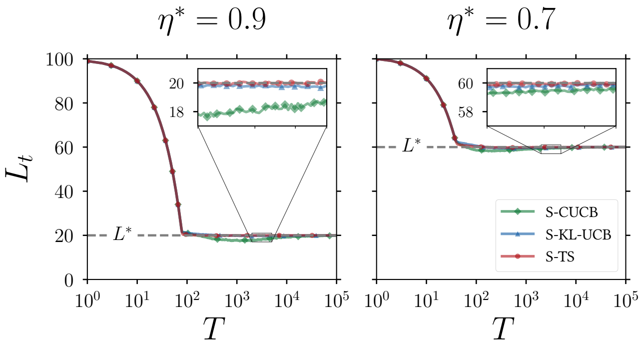
- Maintain $\hat{\mu}_i$ for each arm over a sliding window of adaptive length



$\rightarrow$ Scaling Thompson Sampling with ADWIN (S-TS-A)

Evaluation – Synthetic (Static)

Static: $K = 100$ arms, $T = 10^5$, and μ_i distributed linearly in [0, 1]

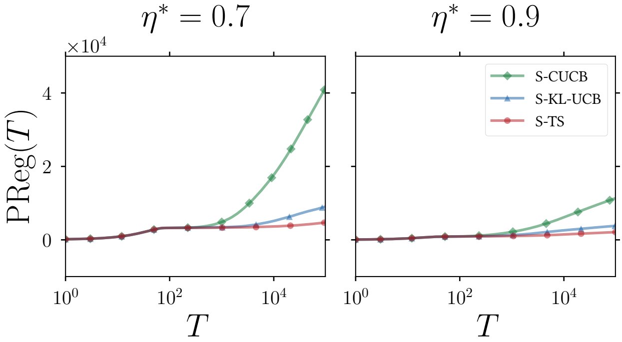


→ Scaling Bandits converge to optimal number of plays L^*

→ S-TS has the lowest regret

Evaluation – Synthetic (Static)

Static: $K = 100$ arms, $T = 10^5$, and μ_i distributed linearly in $[0, 1]$

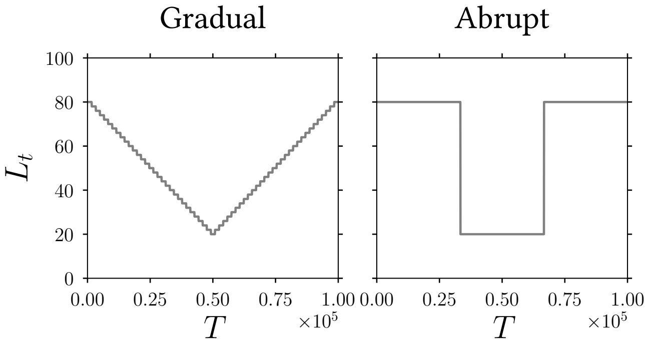


→ Scaling Bandits converge to optimal number of plays L^*

→ S-TS has the lowest regret

Evaluation – Synthetic (Non-Static)

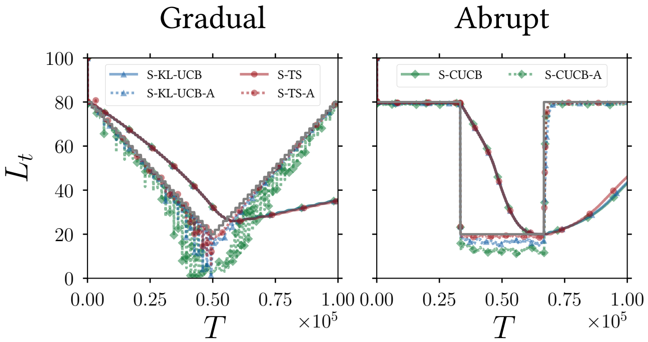
Non-Static: $K = 100$ arms, $T = 10^5$, "gradual" and "abrupt" changes



$\rightarrow$ Bandits based on ADWIN can adapt to gradual and abrupt changes
$\rightarrow$ S-TS with ADWIN has the lowest regret

Evaluation – Synthetic (Non-Static)

Non-Static: $K = 100$ arms, $T = 10^5$, "gradual" and "abrupt" changes

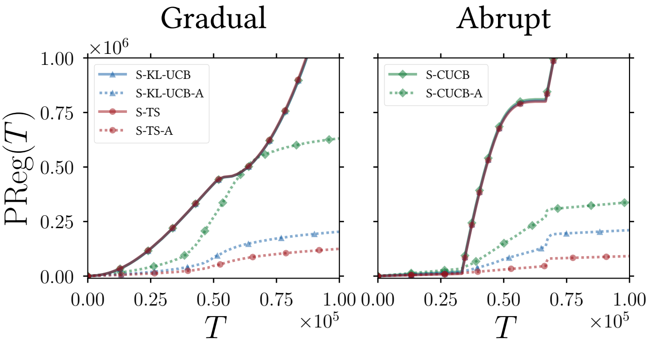


→ Bandits based on ADWIN can adapt to gradual and abrupt changes

→ S-TS with ADWIN has the lowest regret

Evaluation – Synthetic (Non-Static)

Non-Static: $K = 100$ arms, $T = 10^5$, “gradual” and “abrupt” changes

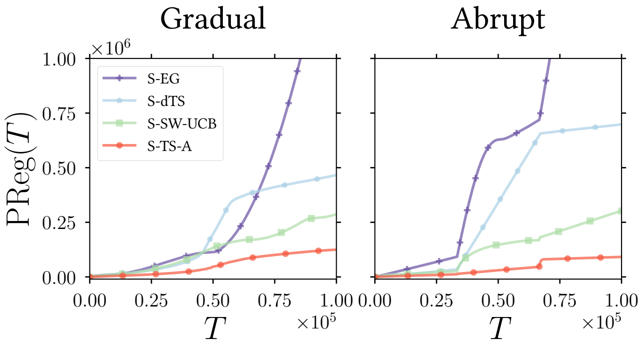


→ Bandits based on ADWIN can adapt to gradual and abrupt changes

→ S-TS with ADWIN has the lowest regret

Evaluation – Synthetic (Non-Static)

Non-Static: $K = 100$ arms, $T = 10^5$, "gradual" and "abrupt" changes



Comparison: ϵ-Greedy (Sutton and Barto, 1998), discounted TS (dTS) (Raj and Kalyani, 2017), Sliding Window UCB (SW-UCB) (Garivier and Moulines, 2008)

Evaluation – Real-world

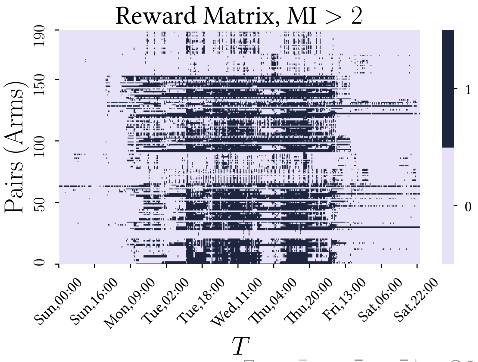
Bioliq power plant – 20 sensors, 1 week monitoring

- Mutual Information (MI) over sliding window
  - Window Size: 1000 points ($\sim$ 15 minutes)
  - Step size: 100 points

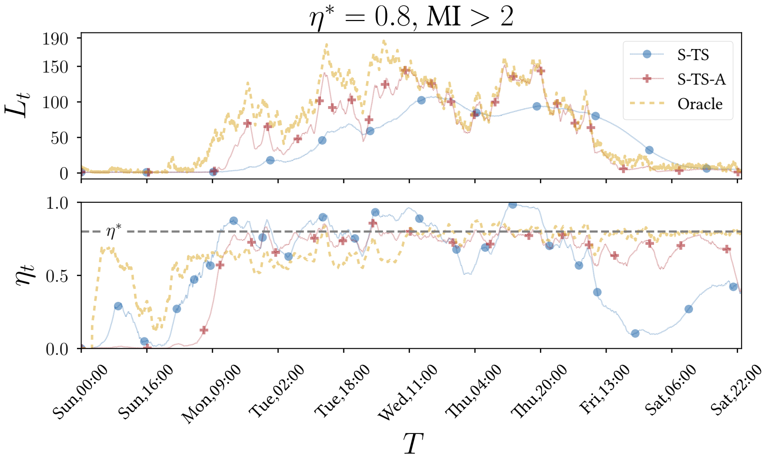
Bandit as a "monitoring system"

- If MI ≥ 2, Reward $= 1$

$\rightarrow K = 190$ arms, $T = 6048$

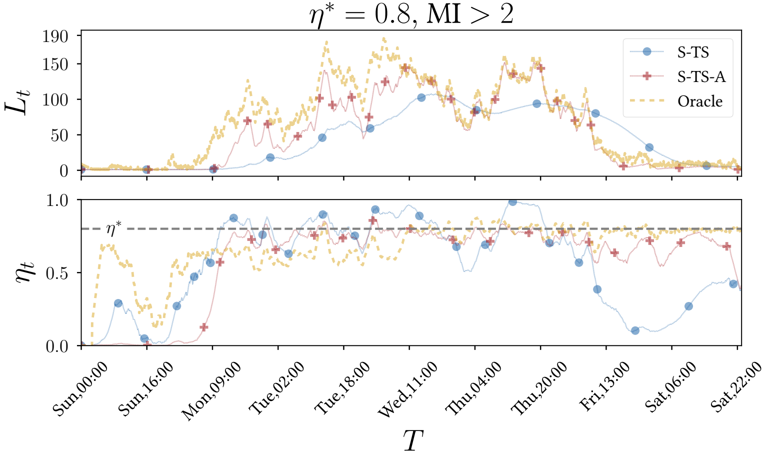


Evaluation – Real-world



→ Scaling of S-TS with ADWIN follows the scaling of the Oracle.

Evaluation – Real-world



→ Scaling of S-TS with ADWIN follows the scaling of the Oracle.

Conclusion

Karlsruhe Institute of Technology

- Scaling Multi-Armed Bandit (S-MAB):
  - Leverage the Multiple-Play MAB with a “scaling policy”
- Theoretical guarantee: Logarithmic regret and “pull regret”
- We combine S-MAB with ADWIN (Bifet and Gavaldà, 2007)
  - Handle the non-static setting
- Evaluation against a real-world use case
  - State-of-the-art performance
  - Code & data: `https://github.com/edouardfouche/S-MAB`

References I

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (1995). Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331.

Bifet, A. and Gavaldà, R. (2007). Learning from time-changing data with adaptive windowing. In *SDM*, pages 443–448. SIAM.

Chen, W., Hu, W., Li, F., Li, J., Liu, Y., and Lu, P. (2016). Combinatorial multi-armed bandit with general reward functions. In *NIPS*, pages 1651–1659.

Garivier, A. and Cappé, O. (2011). The KL-UCB algorithm for bounded stochastic bandits and beyond. In *COLT*, volume 19 of *JMLR Proceedings*, pages 359–376. JMLR.org.

Garivier, A. and Moulines, E. (2008). On Upper-Confidence Bound Policies for Non-Stationary Bandit Problems. *CoRR*, abs/0805.3415.

Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson sampling: An asymptotically optimal finite-time analysis. In *ALT*, volume 7568 of *Lecture Notes in Computer Science*, pages 199–213. Springer.

References II

Komiyama, J., Honda, J., and Nakagawa, H. (2015). Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1152–1161. JMLR.org.

Raj, V. and Kalyani, S. (2017). Taming non-stationary bandits: A bayesian approach. *CoRR*, abs/1707.09727.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning : an introduction*. Adaptive computation and machine learning. MIT Press.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

Uchiya, T., Nakamura, A., and Kudo, M. (2010). Algorithms for adversarial bandit problems with multiple plays. In *ALT*, volume 6331, pages 375–389. Springer.