Karlsruhe Institute of Technology

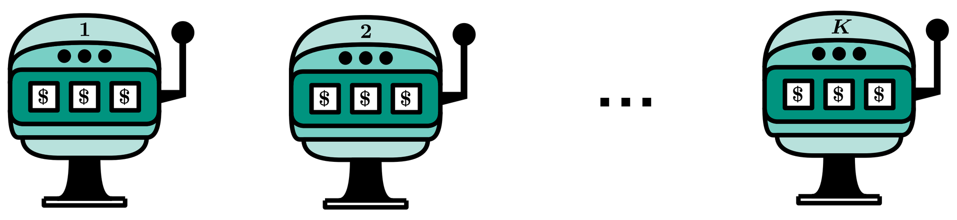
Institute for Program Structures and Data Organization (IPD)
Department of Informatics
& The University of Tokyo – Institute of Industrial Science

Scaling Multi-Armed Bandit Algorithms

Edouard Fouché*, Junpei Komiyama** and Klemens Böhm*
*Karlsruhe Institute of Technology (KIT), **The University of Tokyo

Motivation

The MAB is a fundamental model for sequential decision-making…



- Let there be a set of K arms, $[K] = \{1, 2, \dots, K\}$
  - $i \in [K]$ is associated to a distribution $\mathcal{B}(\mu_i)$ with unknown μ_i
- At each round $t = 1, 2, \dots, T$:
  - **1.** The forecaster chooses **one** arm $i \in [K]$
  - **2.** She observes a reward $X_t \sim \mathcal{B}(\mu_i)$
  - **3.** She updates her estimation $\hat{\mu}_i$ of μ_i
- The goal of the forecaster is to maximize her gain, i.e., $\sum_{t=1}^{T} X_t$

Extension: The Multiple-Play MAB (MP-MAB) [6, 9]

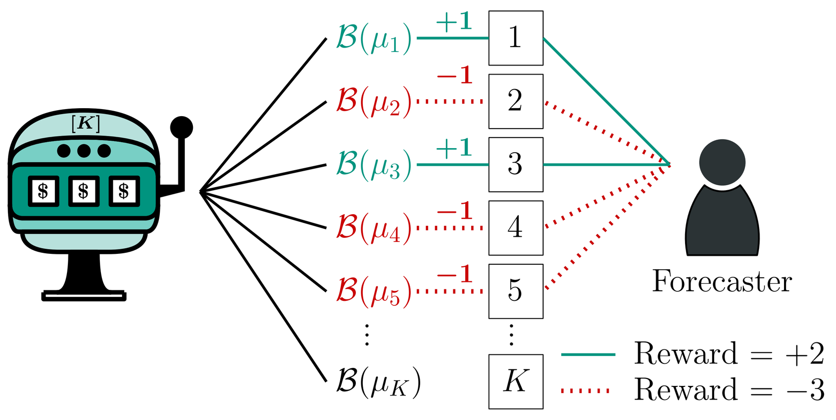
- The forecaster chooses $1 \leq L \leq K$ arm(s) per round

But it has several limitations:

- Typically, playing an arm is associated to a cost
  - L is fixed, and an "efficient" number of plays is unknown
    - Playing too many arms leads to negative gain
    - Playing not enough arms is a loss of potential gain
- Also, the distribution parameters $\mu_1, \dots, \mu_K$ may vary over time

Challenges:

- **C1:** Top-arms Identification
- **C2:** Scale Decision
- **C3:** Change Adaptation



- We propose the **Scaling Multi-Armed Bandit (S-MAB)**
  - The forecaster adapts the number of plays → **C1**, **C2**, **C3**

Real-world Applications

Online Advertisement

Financial Investments

Data Stream Monitoring

Our Use Case: Correlation Monitoring

- Bioliq®: Experimental pyrolysis plant at KIT
  - https://bioliq.de
- S-MAB on Bioliq data stream:
  - Arms ↔ sensor pairs
  - Reward: $MI \geq \Gamma$?

- We release our source code and data:

https://github.com/edouardfouche/S-MAB

Our Contributions

1. We leverage the MP-MAB [6, 9] by introducing a so-called "scaling policy"
   - We prove that the policy converges to an "optimal" number of plays
   - → S-MAB has logarithmic regret and logarithmic "pull regret"
2. We combine S-MAB with ADWIN [1] for the non-static setting
   - ADWIN maintains estimates of $\mu_1, \dots, \mu_K$, which are changing over time
   - S-MAB with ADWIN can handle both gradual and abrupt changes
3. We evaluate against synthetic and real-world data
   - S-MAB shows excellent performance compared to the state of the art

Our Approach & Evaluation

Problem Definition: MP-MAB with efficiency constraint

- $I_t \subset [K]$ is 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

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

where η^* is a user-/application-specific efficiency threshold

- e.g., if reward ≤ 1 and cost $= 1$ for each arm, then $\eta^* > 0.5$

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

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

General Scaling Multi-Armed Bandit

Require: Set of arms $[K]$, target efficiency η^*

1: $\hat{\mu}_1, \dots, \hat{\mu}_K \leftarrow 1$
2: $L_1 \leftarrow K$
3: **for** $t = 1, 2, \dots, T$ **do**
4: $I_t \leftarrow \text{CHOOSE}([K], L_t, \hat{\mu}_1, \dots, \hat{\mu}_K)$
5: $X_t \leftarrow \text{OBSERVE}(I_t)$
6: $\hat{\mu}_1, \dots, \hat{\mu}_K \leftarrow \text{UPDATE}(X_t, \hat{\mu}_1, \dots, \hat{\mu}_K)$
7: $L_{t+1} \leftarrow \text{SCALE}(L_t, \eta^*, \hat{\mu}_1, \dots, \hat{\mu}_K)$

Any bandit algorithm, e.g.:

- Thompson Sampling (TS) [6]
- UCB-type [2]
- Exp3 [9]
- …

→ Our extension is independent from the underlying "base bandit"

The standard regret Reg and the "pull regret" $PReg$ (static):

$$Reg(T) = \sum_{t=1}^{T} \left[\max_{\mathcal{J} \subseteq [K], |\mathcal{J}| = L_t} \sum_{j \in \mathcal{J}} \mu_j - \sum_{i \in I_t} \mu_i \right] \qquad PReg(T) = \sum_{t=1}^{T} |L^* - L_t|$$

Scaling Policy: Kullback-Leibler Scaling (KL-S)

$$L_{t+1} = \begin{cases} L_t - 1 & if\ \hat{\eta}_t \leq \eta^* \\ L_t + 1 & if\ \hat{\eta}_t > \eta^*\ and\ \hat{B}_t > \eta^* \\ L_t & otherwise \end{cases} \qquad \hat{\eta}_t = \frac{1}{L_t} \sum_{i}^{I_t} \hat{\mu}_i$$

$$\hat{B}_t = \frac{L_t}{L_t + 1} \hat{\eta}_t + \frac{1}{L_t + 1} b_{\widehat{L_t+1}}$$

- $b_i(t)$ is the KL-UCB index of arm i [4]
- $\widehat{L_t+1}$ is the arm with the (L_t+1)-th largest index

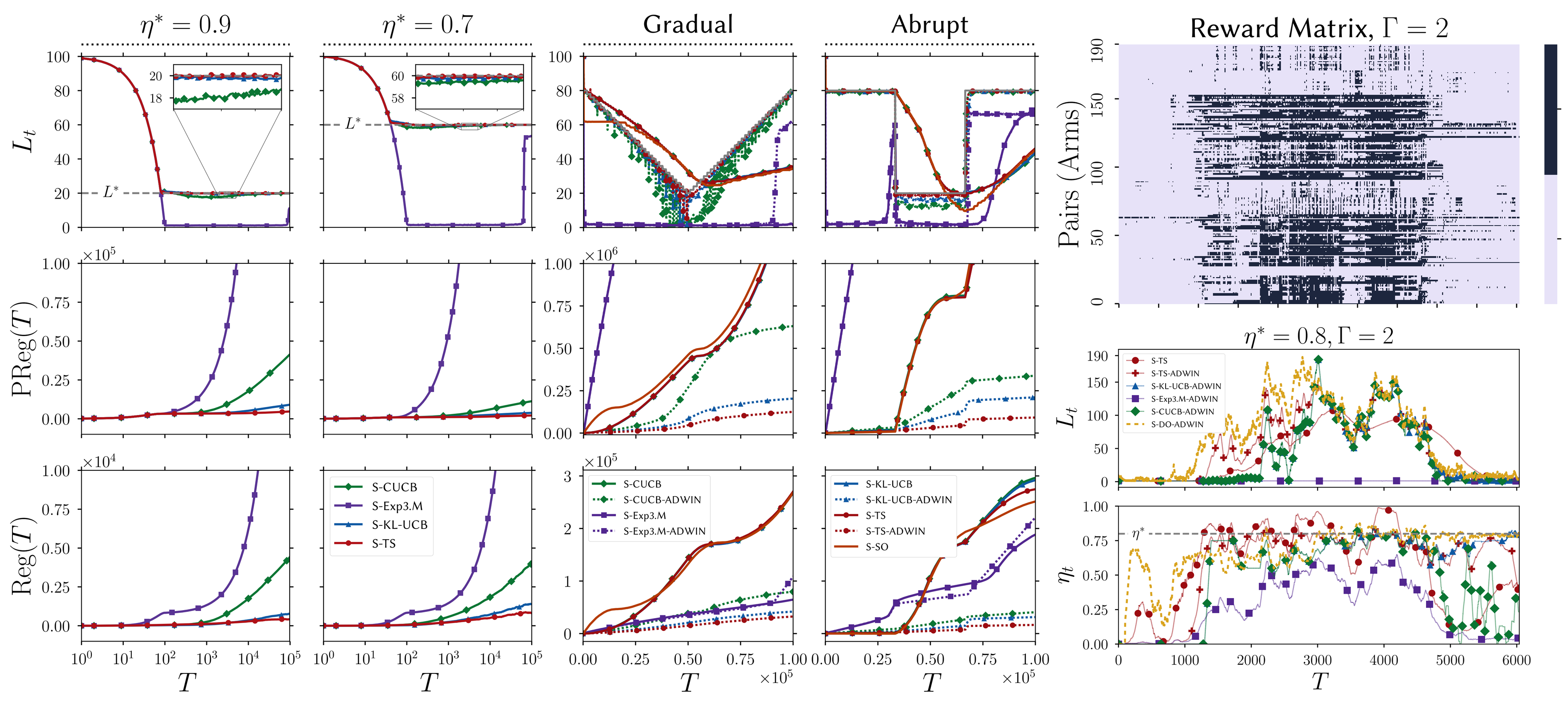
Theorem (Logarithmic regret and logarithmic "pull regret")

« The S-MAB has logarithmic regret and logarithmic pull regret, with respect to increasing time T, for any "base bandit" with logarithmic regret » (see proof in [3])

→ S-TS and S-KL-UCB and S-UCB have logarithmic regret and pull regret

Experiments: We publish our benchmark data sets (see our **GitHub repository**)

- Static
- Non-static
- Real-world



References

[1] Bifet, A. and Gavaldà, R. (2007). Learning from Time-Changing Data with Adaptive Windowing. In SDM '19. SIAM

[2] Chen, W., Hu, W., Li, F., Li, J., Liu, Y., and Lu, P. (2016). Combinatorial Multi-Armed Bandit with General Reward Functions. In NIPS '16.

[3] Fouché, E., Komiyama, J. and Böhm, K. (2019). Scaling Multi-Armed Bandit Algorithms. In KDD '19.

[4] Garivier, A. and Cappé, O. (2011). The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond. In COLT '11.

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

[6] 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 '15.

[7] Raj, V. and Kalyani, S. (2017). Taming Non-stationary Bandits: A Bayesian Approach. CoRR, abs/1707.09727.

[8] 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.

[9] Uchiya, T., Nakamura, A., and Kudo, M. (2010). Algorithms for Adversarial Bandit Problems with Multiple Plays. In ALT '10.