Memory Bandits: a Bayesian approach for the Switching Bandit Problem

Réda ALAMI[1,3], Odalric-Ambrym MAILLARD[2], Raphael FERAUD[3]
[1]Paris-Saclay University, [2]Inria Lille, [3]Orange Labs

BayesOpt 2017

Multi-Armed Bandit

For each step $t = 1, ..., T$

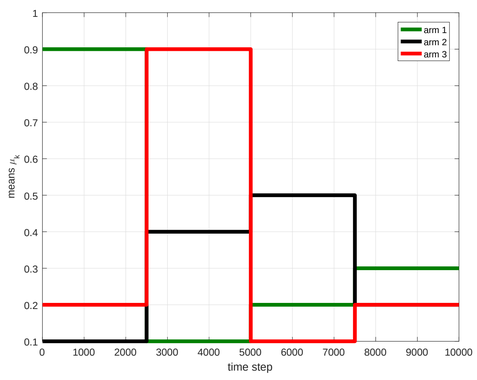
- The player chooses an arm $k_t \in \mathcal{K}$
- The reward k_t is revealed $x_{k_t} \in [0, 1]$
- Bernoulli rewards: $x_{k_t} \sim \mathcal{B}(\mu_{k_t,t})$

Minimize the pseudo regret:
$\mathcal{R}(T) = \sum_{t=1}^{T} \mu_t^\star - \mathbb{E}\left[\sum_{t=1}^{T} x_{k_t}\right]$,
where $\mu_t^\star = \max_k \mu_{k,t}$.

Switching Environment

$$\mu_{k,t} = \begin{cases} \mu_{k,t-1} & \text{probability } 1-\rho \\ \mu_{new} \sim \mathcal{U}(0,1) & \text{probability } \rho \end{cases}$$

where ρ is the switching rate.



Two kinds of switch

Global switch: all arms change their means.
Per-arm switch: an arm changes its mean.

Thompson Sampling (TS)

$$\begin{cases} \text{success counter} : \alpha_k = \#(x_k = 1) + \alpha_0 \\ \text{failure counter} : \beta_k = \#(x_k = 0) + \beta_0 \end{cases}$$
At each $t \leq T$:

- *Characterization*: $\theta_k \sim \text{Beta}(\alpha_k, \beta_k)$
- *Decision*: $k_t = \arg\max_k \theta_k$
- *Update*: $\begin{cases} \alpha_k = \alpha_k + 1 & \text{if } x_{k_t} = 1 \\ \beta_k = \beta_k + 1 & \text{if } x_{k_t} = 0 \end{cases}$

$\mathcal{R}(T) \leq (1+\epsilon) \sum_{\mu_k \neq \mu^*} \frac{\Delta_k (\log(T) + \log\log(T))}{KL(\mu_k, \mu^*)}$ (Lai and Robbins (1985) lower bound)

References

R. P. Adams and D.J.C MacKay, *Bayesian online changepoint detection*, arXiv, 2007.

J. Mellor and J. Shapiro, *Thompson Sampling in switching environments with Bayesian online changepoint detection*, AISTATS, 2013.

Global Switching TS with Bayesian Aggregation

Growing number of Thompson Sampling $f_{i,t}$: i denotes the starting time and t the current time. Let $\mathbb{P}(f_{i,t})$ be the probability at time t of the Thompson sampling starting at time i.

Initialization: $\mathbb{P}(f_{1,1}) = 1, t = 1$,
$\forall k \in \mathcal{K}\ \alpha_{k,f_{1,1}} = \alpha_0,\ \beta_{k,f_{1,1}} = \beta_0$

1 Decision process: at each time t:

- $\forall\, i < t, \forall k$: $\theta_{k,f_{i,t}} \sim \text{Beta}\left(\alpha_{k,f_{i,t}}, \beta_{k,f_{i,t}}\right)$
- Play (**Bayesian Aggregation**):

$$k_t = \arg\max_k \sum_{i<t} \mathbb{P}(f_{i,t})\, \theta_{k,f_{i,t}}$$

2 Instantaneous gain update:

$$\forall\, i < t\ \mathbb{P}(x_t | f_{i,t}) = \begin{cases} \frac{\alpha_{k,f_{i,t}}}{\beta_{k,f_{i,t}} + \alpha_{k,f_{i,t}}} & \text{if } x_{k_t} = 1 \\ \frac{\beta_{k,f_{i,t}}}{\beta_{k,f_{i,t}} + \alpha_{k,f_{i,t}}} & \text{if } x_{k_t} = 0 \end{cases}$$

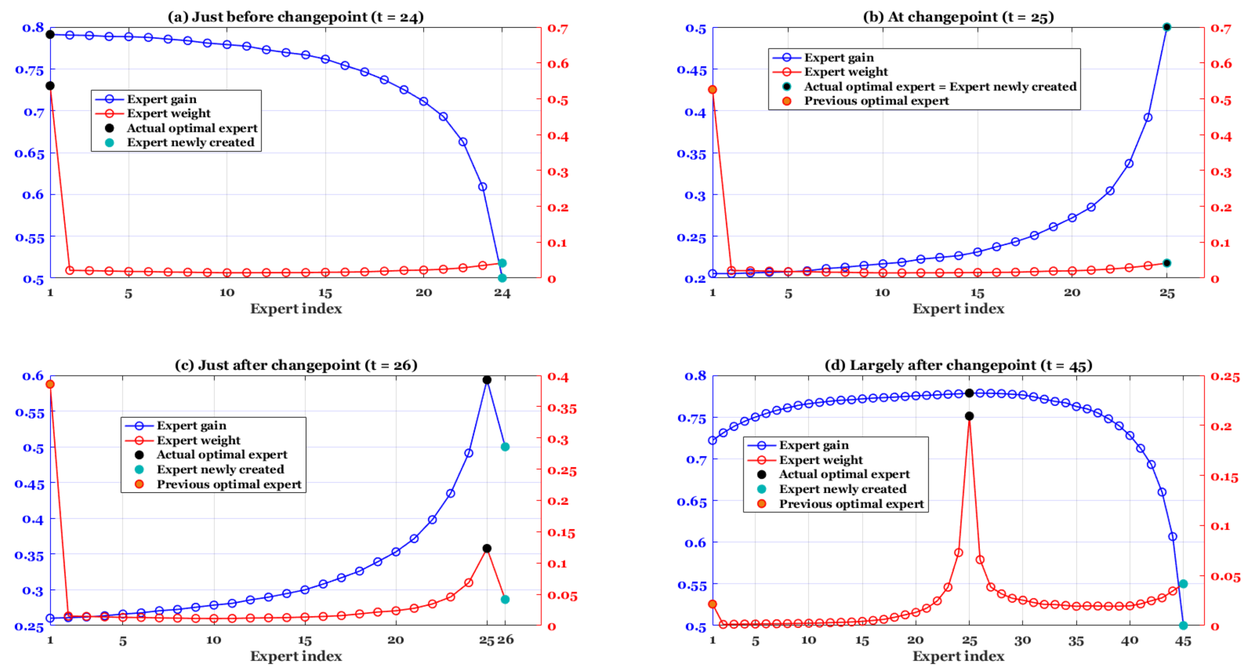
3 Arm hyperparameters update:

$$\forall\, i < t \begin{cases} \alpha_{k,f_{i,t}} = \alpha_{k,f_{i,t}} + 1 & \text{if } x_{k_t} = 1 \\ \beta_{k,f_{i,t}} = \beta_{k,f_{i,t}} + 1 & \text{if } x_{k_t} = 0 \end{cases}$$

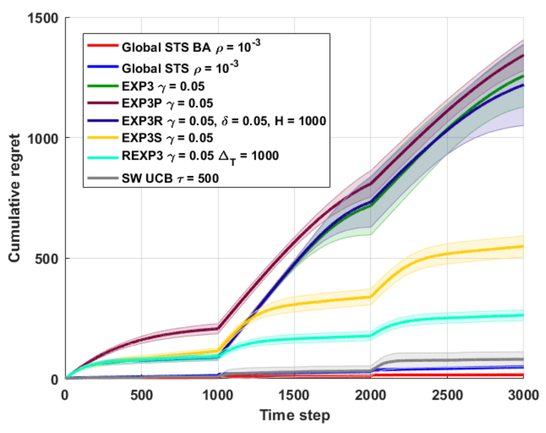
4 Distribution of experts update:

- Update: $\forall i < t, \mathbb{P}(f_{i,t}) \propto (1-\rho) \cdot \mathbb{P}(x_t | f_{i,t-1}) \cdot \mathbb{P}(f_{i,t-1})$
- Create $f_{t,t}$: $\mathbb{P}(f_{t,t}) \propto \rho \sum_{i=1}^{t-1} \mathbb{P}(x_t | f_{i,t-1}) \cdot \mathbb{P}(f_{i,t-1})$
- Prior: $\alpha_{k,f_{t,t}} = \alpha_0, \beta_{k,f_{t,t}} = \beta_0$

Tracking the best expert



Comparison with state-of-the-art



Sensitivity analysis of parameters