MEMORY BANDITS: TOWARDS THE SWITCHING BANDIT PROBLEM BEST RESOLUTION

RÉDA ALAMI[1,3], ODALRIC-AMBRYM MAILLARD[2], RAPHAEL FERAUD[3]
[1] INRIA-SACLAY (LRI), [2]INRIA LILLE (SEQUEL), [3]ORANGE LABS

MULTI-ARMED BANDIT

For each step $t = 1, \dots, T$

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

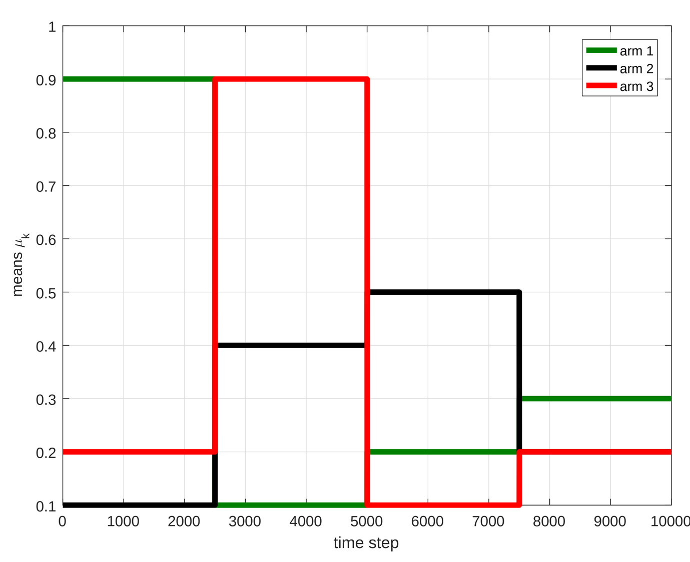
Objective: Minimize the pseudo regret $\mathcal{R}_T$:

$$\mathcal{R}_T = \underbrace{\sum_{t=1}^{T} \mu_t^\star}_{\text{Best policy}} - \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} x_{k_t}\right]}_{\text{Your policy}} \quad \mu_t^\star = \max_k \mu_{k,t}$$

SWITCHING ENVIRONMENT

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

where ρ is the switching rate.



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 step $t = 1, \dots, T$:

1. *Characterization*: $\theta_k \sim \text{Beta}(\alpha_k, \beta_k)$
2. *Decision*: $k_t = \arg\max_k \theta_k$
3. *Play*: $x_{k_t} \sim \mathcal{B}(\mu_{k_t})$
4. *Update*: $\begin{cases} \alpha_{k_t} = \alpha_{k_t} + 1 & \text{if } x_{k_t} = 1 \\ \beta_{k_t} = \beta_{k_t} + 1 & \text{if } x_{k_t} = 0 \end{cases}$

$\mathcal{R}_T \leq (1+\epsilon) \sum_k \frac{\mu^\star - \mu_k}{KL(\mu_k, \mu^\star)} (\log T + \log\log T)$
(Lai and Robbins (1985) lower bound)
$KL(\bullet, \bullet)$ = Kullback-Leibler divergence

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

Learning with a growing number of Thompson Sampling $f_{i,t}$: i denotes the starting time and t the current time. $\mathbb{P}(f_{i,t})$: weight 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:

- $\forall\, i \leq t, \forall k$: $\theta_{k,f_{i,t}} \sim \text{Beta}(\alpha_{k,f_{i,t}}, \beta_{k,f_{i,t}})$
- 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:

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

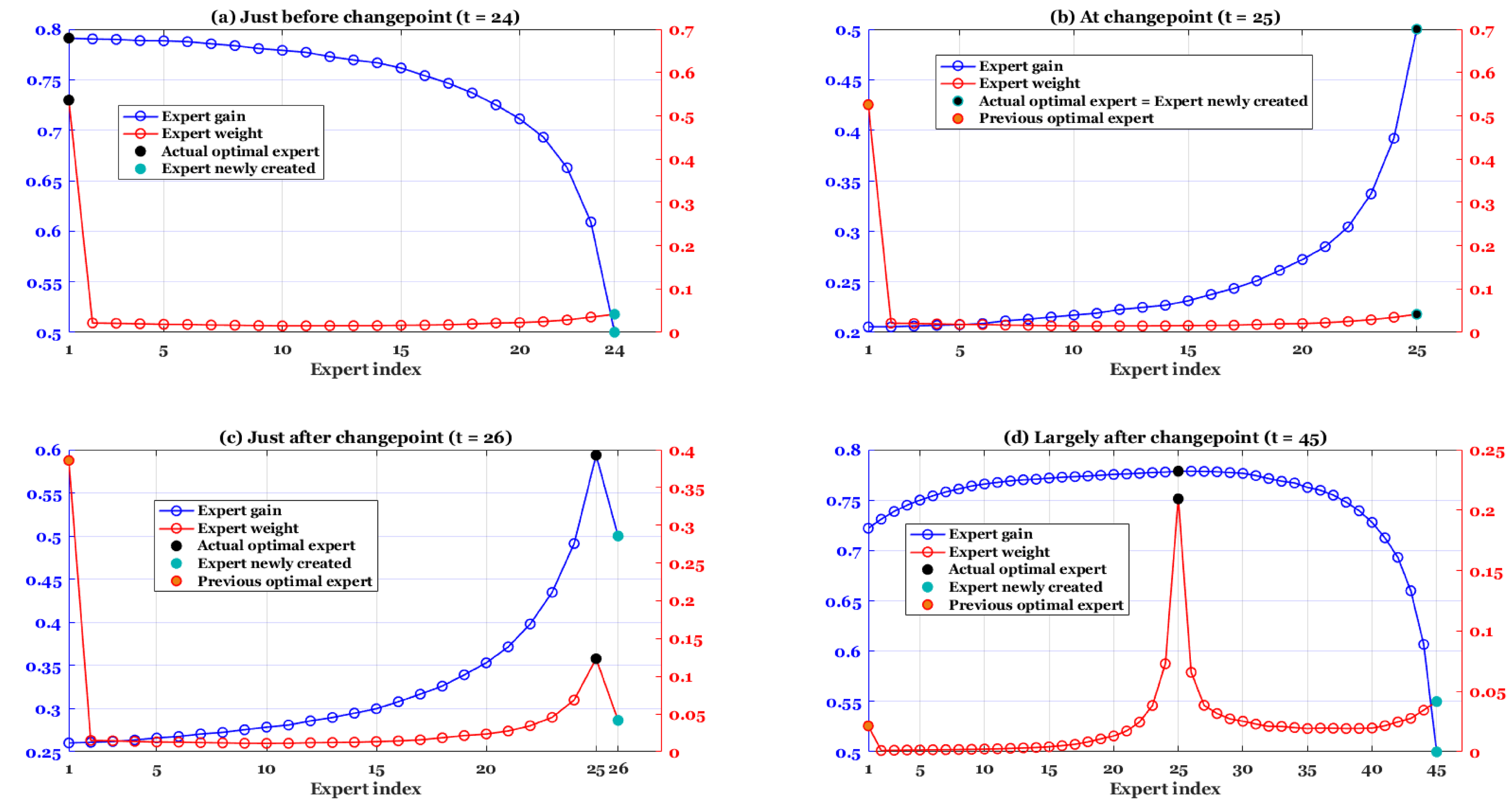
-3- Arm hyperparameters update:

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

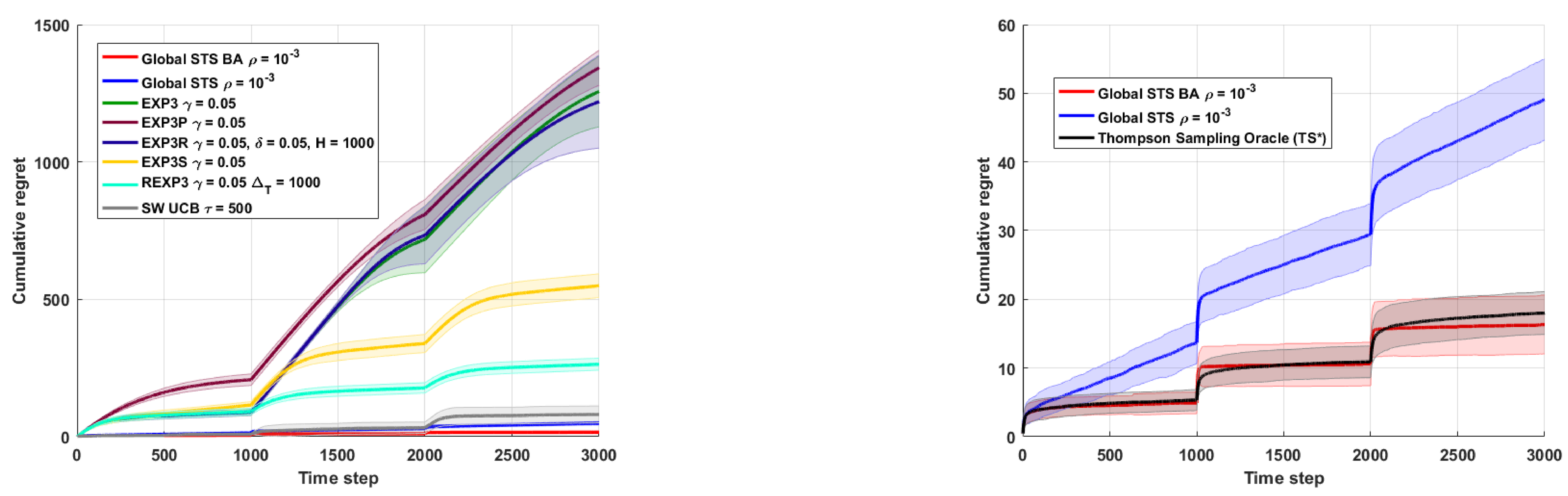
-4- Distribution of experts update:

- Update previous experts: $\mathbb{P}(f_{i,t+1}) \propto (1-\rho) \cdot \mathbb{P}(\mathbf{x_t} | \mathbf{f_{i,t}}) \cdot \mathbb{P}(f_{i,t}) \quad \forall i \leq t$
- Create new expert $f_{t+1,t+1}$: $\mathbb{P}(f_{t+1,t+1}) \propto \rho \sum_{i=1}^{t} \mathbb{P}(f_{i,t})$
- Prior: $\alpha_{k,f_{t,t}} = \alpha_0, \beta_{k,f_{t,t}} = \beta_0$

TRACKING THE OPTIMAL EXPERT



COMPARISON WITH STATE-OF-THE-ART



SENSITIVITY ANALYSIS OF PARAMETERS (ρ AND M)