Memory Bandits: Towards the Switching Bandit Problem Best Resolution
Réda Alami, Odalric-Ambrym Maillard, Raphaël Féraud

To cite this version:
Réda Alami, Odalric-Ambrym Maillard, Raphaël Féraud. Memory Bandits: Towards the Switching Bandit Problem Best Resolution. MLSS 2018 - Machine Learning Summer School, Aug 2018, Madrid, Spain. <hal-01879251>

HAL Id: hal-01879251
https://hal.archives-ouvertes.fr/hal-01879251
Submitted on 22 Sep 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
MEMORY BANDITS: TOWARDS THE SWITCHING BANDIT PROBLEM BEST RESOLUTION
REDA ALAMI1,3, ODALRIC-AMBRYM MAILLARD2, RAPHAEL FERAUD3
1 INRIA-Saclay (LRI), 2 INRIA Lille (SEQUEL), 3 Orange Labs

MULTI-ARMED BANDIT

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

Objective: Minimize the pseudo regret \( R_T \):

\[
R_T = \sum_{t=1}^{T} \max_{k} \mu_{k,t} - E \left[ \sum_{t=1}^{T} x_{k_t} \right] \]

where \( \rho \) is the switching rate.

GLOBAL SWITCHING TS WITH BAYESIAN AGGREGATION

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

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

- Decision process: at each time \( t \):
  - \( \forall i \leq t, \forall k: \theta_{k,f,i} \sim \text{Beta}(\alpha_{k,f,i}, \beta_{k,f,i}) \)
  - Play (Bayesian Aggregation):

\[
k_t = \arg \max_{k} \sum_{i \leq t} \mathbb{P}(f_{i,i}) \theta_{k,i} \]

- Instantaneous gain update:

\[
\forall i \leq t \mathbb{P}(x_t|f_{t,i}) = \left\{ \begin{array}{ll}
\frac{\alpha_{k,f,i} + x_t}{\alpha_{k,f,i} + \beta_{k,f,i}} & \text{if } x_t = 1 \\
\frac{\beta_{k,f,i} + 1}{\alpha_{k,f,i} + \beta_{k,f,i} + 1} & \text{if } x_t = 0
\end{array} \right.
\]

- Arm hyperparameters update:

\[
\forall i \leq t \left\{ \begin{array}{ll}
\alpha_{k,f_i+1} = \alpha_{k,f,i} + 1 & \text{if } x_t = 1 \\
\beta_{k,f_i+1} = \beta_{k,f,i} + 1 & \text{if } x_t = 0
\end{array} \right.
\]

- Distribution of experts update:

\[
\text{Update previous experts: } \mathbb{P}(f_{t+1,i+1}) \propto (1 - \rho) \cdot \mathbb{P}(x_t|f_{t+1,i}) \cdot \mathbb{P}(f_{t,i}) \\
\text{Create new expert } f_{t+1,i+1}: \mathbb{P}(f_{t+1,i+1}) \propto \rho \sum_{i=1}^{t} \mathbb{P}(f_{t,i})
\]

\[
\text{Prior: } \alpha_{k,f,i} = \alpha_0, \beta_{k,f,i} = \beta_0
\]

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 \( t = 1, \ldots, T \):
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:
   - \( \alpha_k = \alpha_k + 1 \) if \( x_{k_t} = 1 \)
   - \( \beta_k = \beta_k + 1 \) if \( x_{k_t} = 0 \)

\[
R_T \leq (1 + \epsilon) \sum_{k=1}^{K} \frac{\rho}{\beta_k + \epsilon} (\log T + \log \log T)
\]
(Lai and Robbins (1985) lower bound)

\[
KL(\mu, \cdot) = \text{Kullback-Leibler divergence}
\]

SWITCHING ENVIRONMENT

\[
\mu_{k,t} = \begin{cases}
\mu_{k,t-1} \text{ probability } 1 - \rho \\
n_{k,new} \sim U(0, 1) \text{ probability } \rho
\end{cases}
\]

TRACKING THE OPTIMAL EXPERT

COMPARISON WITH STATE-OF-THE-ART

SENSITIVITY ANALYSIS OF PARAMETERS (\( \rho \) AND \( M \))

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