European Workshop on Reinforcement Learning 14 (2018) October 2018, Lille, France.

Thompson Sampling for the non-Stationary Corrupt Multi-Armed Bandit

Réda Alami ALAMI1859@HOTMAIL.FR
Orange Labs
2 Avenue Pierre Marzin
22300, Lannion, France

Abstract

We propose an extension of the corrupt multi-armed bandit problem proposed in Gajane et al. (2018) where the distributions of reward and feedback are non-stationary. We also propose an extension with three variants of the Global Switching Thompson Sampling proposed in Mellor and Shapiro (2013) for the corrupted setting. This extension is based on the aggregation of a growing number of experts seen as learners. Finally, we conduct experiments providing evidences that in practice our proposal compares favorably with the oracle that exactly knows the location of the environment changes.

Keywords: Multi-armed bandit, sequential learning, corrupted feedback, Thompson Sampling.

1. Introduction and motivations

The Multi-Armed Bandit problem formalizes the fundamental exploration-exploitation dilemma that appears in decision making problems facing partial information, where decisions have to be taken over time (discrete turns) and impact both the rewards and the information withdrawn. Specically, a set of K arms (or actions) is available to the decision-maker (player). At each turn, he has to choose one arm and observes a *feedback* corresponding to the played arm, ignoring what the observed *feedback* would have been, if he had played another arm. The player faces the dilemma of exploring, that is playing an arm whose mean reward is loosely estimated in order to build a better estimate, or exploiting, that is playing a seemingly best arm based on current estimates in order to maximize its cumulative reward. The accuracy of the player policy at a given time horizon is typically measured in terms of regret, that is the difference between the cumulative rewards of the player and the one that could have been acquired by a policy assumed to be optimal.

In the classical MAB problem, the feedback is the observation of the reward itself. However, in some pratical settings, this assumption does not hold true. The classical example of such setting is the adaptive routing where positive feedback means that the corresponding path is usable but no feedback could either mean that the corresponding path is unusable or the feedback was dropped due to extraneous issues. This has encouraged the community to introduce what we call the corrupt bandit problem.

The corrupt multi-armed bandit problem has been proposed in Gajane et al. (2018), and has been resolved efficiently by extending the classical Thompson Sampling to the corrupt setting. This algorithm is called *Thompson Sampling with Corrupted Feedback*. However, in such these settings, the distribution of rewards and feedbacks are assumed to

be stationary. In this paper, we propose an extension of the corrupt bandit to the abrupt switching environment. Moreover, we propose an adaptation of Thompson Sampling with Corrupted Feedback to the switching environment.

2. Problem formulation

2.1 The Non-stationary Corrupt Multi-Armed Bandit Problem

Let us consider an agent facing a non-stationary stochastic corrupt multi-armed bandit problem characterized by a set $\mathcal{K} = \{1, ..., K\}$ of K independent arms. At each round $t \in [\![1, T]\!]$, the agent chooses to observe one of the K possible actions. When playing the arm k_t at time t, two events take place. First, a reward x_t is received, where $x_t \sim \mathcal{B}(\mu_{k_t,t})$ is a random variable drawn from a Bernoulli distribution of expectation $\mu_{k,t}$. Then, a feedback y_t is observed, where $y_t \sim \mathcal{B}(\nu_{k_t,t})$ is a random variable drawn from a Bernoulli distribution of expectation $\nu_{k,t}$. In such that setting, we assume the existence of a loose link between the mean reward $\mu_{k,t}$ and the mean feedback $\nu_{k,t}$ of arm k at time t via a known $[0, 1]$-monotonic continuous function g_k which is called *mean-corruption function* such that: $g_k(\mu_{k,t}) = \nu_{k,t}$. It should be noted that g_k may be completely different from an arm to another. Thus, let $\mathcal{G} = \{g_k : k \in \mathcal{K}\}$ denote the set of mean-corruption function available to the agent. Finally, let $\mu_t^\star = \max_{k\in\mathcal{K}} \{\mu_{k,t}\}$ denotes the best expected reward at round t, $k_t^\star = \arg\max_{k\in\mathcal{K}} \{\mu_{k,t}\}$ the best arm at round t, k_t the action chosen by the decision-maker at time t and y_t the feedback observed at the same time.

2.2 Piece-wise stationary Bernoulli distributions

We assume that there exists a parameter $\rho \in (0, 1)$ such that both $\mu_{k,t}$ and $\nu_{k,t}$, i.e. the reward mean and the feedback mean of arm k at time t follow a global switching model:

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

When the environment is modeled by eq (1) for all $k \in \mathcal{K}$, the problem setting is called a *Global Switching Corrupt Multi-Armed Bandit* (**GS-CMAB**), i.e. when a switch occurs *all* arms change their expected rewards and expected feedbacks. There exists a more general setting where changes occur independently for each arm k (i.e. arms change points are independent from an arm to another). In this case, the problem setting is called a *Per-arm Switching Corrupt Multi-Armed Bandit*. In this paper, we will focus more on the first setting (GS-CMAB).

2.3 Sequence of change points

It should be noted that for each GS-CMAB, there exists a non-decreasing change points sequence of length Γ_T denoted by $(\tau_\kappa)_{\kappa\in[\![1,\Gamma_T+1]\!]} \in \mathbb{N}^{\Gamma_T+1}$ where:

$$\begin{cases} \forall\ \kappa \in [\![1, \Gamma_T]\!],\ \forall\ t \in \mathcal{T}_\kappa = [\![\tau_\kappa + 1, \tau_{\kappa+1}]\!],\ \forall\ k \in \mathcal{K},\ \ \mu_{k,t} = \mu_{k,[\kappa]} \text{ and } \nu_{k,t} = \nu_{k,[\kappa]} \\ \tau_1 = 0 < \tau_2 < ... < \tau_{\Gamma_T+1} = T \end{cases}$$

In this case, $\mu_{[\kappa]}^\star = \max_k \left\{\mu_{k,[\kappa]}\right\}$ denotes the highest expected reward at epoch $\mathcal{T}_\kappa$.

2.4 Cumulative regret in the piece-wise stationary corrupted environment

In a piece-wise stationary corrupt environment, the pseudo cumulative regret $\mathcal{R}(T)$ up to time T is defined as the expectation of the cumulative difference between the rewards obtained by the *oracle* who plays the best arm $k^{\star}_{[\kappa]}$ at each epoch $\mathcal{T}_{\kappa}$ and those received by our policy such as:

$$\mathcal{R}(T) = \sum_{t=1}^{T} \mu_t^{\star} - \sum_{t=1}^{T} \mu_{k_t,t} = \sum_{\kappa=1}^{\Gamma_T} \left(|\mathcal{T}_{\kappa}| \, \mu^{\star}_{[\kappa]} - \sum_{t=\tau_{\kappa}+1}^{\tau_{\kappa+1}} \mu_{k_t,t} \right)$$

2.5 Thompson Sampling with Corrupted Feedback (TS-CF)

Thompson Sampling with Corrupted Feedback belongs to the Bayesian online learning family. It is leveraging Bayesian tools by maintaining a Beta posterior distribution $\pi_{k,t} = \text{Beta}(\alpha_{k,t}, \beta_{k,t})$ on the feedback distribution of each arm k. Based on the feedback observed y_t, the posterior distribution $\pi_{k,t}$ is updated such as:

$$\pi_{k,t} = \text{Beta}\left(\alpha_{k,t} = \#(\text{feedback} = 1) + \alpha_0, \beta_{k,t} = \#(\text{feedback} = 0) + \beta_0\right)$$

At each time, the agent takes a sample $\theta_{k,t}$ from each $\pi_{k,t}$ and then plays the arm $k_t = \arg\max_k \left\{ \vartheta_{k,t} = g_k^{-1}(\theta_{k,t}) \right\}$. Formally, by denoting $D_{t-1} = \bigcup_{i=1}^{t-1} y_i$ the history of past feedbacks we write: $\theta_t = (\theta_{1,t}, ..., \theta_{K,t}) \sim \mathbb{P}(\theta_t | D_{t-1}) = \prod_{k=1}^{K} \pi_{k,t}$.

Recently, in Gajane et al. (2018) TS-CF has been shown to be asymptotically optimal. Indeed, its pseudo-cumulative regret reaches the lower bound on the pseudo-regret for MAB with Corrupted Feedback.

3. Global Switching Thompson Sampling with Corrupted Feedback (Global-STS-CF)

3.1 Best achievable performance: TS-CF oracle

Let TS-CF* denotes the oracle that knows exactly the location of the change points τ_{κ}. It simply restarts a TS-CF at these change points. Assuming that Γ_T is the overall number of change points observed until the horizon T, then TS-CF* runs successively Γ_T TS-CF procedures starting at $\tau_{\kappa} + 1$ and ending at $\tau_{\kappa+1}$. Since TS-CF is asymptotically optimal (Gajane et al. (2018)), TS-CF* represents the best achievable performance in the global switching corrupted MAB.

3.2 Notion of expert

Let $t \in \mathbb{N}^{\star}$ and $i \in [\![1, t]\!]$. An expert $f_{i,t}$ is a TS-CF procedure which has started at time i. The expert $f_{i,t}$ observes exactly $t - i$ feedbacks from the environment.

3.3 Decision-making based on a growing number of experts

Like in Alami et al. (2017); Mellor and Shapiro (2013), in order to detect the occurrence of the changepoints, we connect the well-known Bayesian online changepoint detector of Adams and MacKay (2007) with the corrupted version of the multi-armed bandit setting. Indeed, at each time step t, a new expert is introduced. One can see the expert $f_{i,t}$ as an index used to get access to the memory saving the hyperparameters of the model created at time t. Thus, dealing with the expert distribution $w_{i,t} = \mathbb{P}(f_{i,t} | D_{t-1})$, the computation of the posterior distribution $\mathbb{P}(\theta_t | D_{t-1})$ takes the following form:

$$\mathbb{P}(\theta_t|D_{t-1}) = \sum_{i=1}^{t} \mathbb{P}(\theta_t|D_{t-1}, f_{i,t})\mathbb{P}(f_{i,t}|D_{t-1}) \tag{2}$$

Building the expert distribution According to the work of Adams and MacKay (2007), the computation of the expert distribution is done recursively such that:

$$\underbrace{\mathbb{P}\left(f_{i,t}|D_{t-1}\right)}_{\texttt{Expert distribution at } t} \propto \sum_{i=1}^{t-1} \overbrace{\mathbb{P}(f_{i,t}|f_{i,t-1})}^{\texttt{change point prior}} \underbrace{\mathbb{P}(y_t|f_{i,t-1}, D_{t-2})}_{\texttt{Instantaneous gain}} \overbrace{\mathbb{P}(f_{i,t-1}|D_{t-2})}^{\texttt{Expert distribution at } t-1} \tag{3}$$

The change point prior $\mathbb{P}(f_{i,t}|f_{i,t-1})$ is naturally computed following eq (1):

$$\mathbb{P}(f_{i,t}|f_{i,t-1}) = (1-\rho)\,\mathbb{1}\left(i < t\right) + \rho\mathbb{1}\left(i = t\right) \tag{4}$$

Thus, the inference model takes the following form (Up to a normalization factor):

$$\begin{cases} \textit{Growth probability:} & \mathbb{P}(f_{i,t}|D_{t-1}) \propto (1-\rho)\cdot\mathbb{P}(y_t|f_{i,t-1}, D_{t-2})\cdot\mathbb{P}(f_{i,t-1}|D_{t-2}) \\ \textit{Change point probability:} & \mathbb{P}(f_{t,t}|D_{t-1}) \propto \rho\sum_{i=1}^{t-1}\mathbb{P}(y_t|f_{i,t-1}, D_{t-2})\cdot\mathbb{P}(f_{i,t-1}|D_{t-2}) \end{cases} \tag{5}$$

Where $\mathbb{P}(y_t|f_{i,t-1}, D_{t-2})$ corresponds to the likelihood of the Bernoulli distribution parametrized with $\frac{\alpha_{k_t,i,t-1}}{\alpha_{k_t,i,t-1}+\beta_{k_t,i,t-1}}$, where $\alpha_{k_t,i,t-1}$ and $\beta_{k_t,i,t-1}$ are the hyper-parameters of the arm k_t learned by the expert $f_{i,t-1}$. Let $l_{i,t-1}$ denotes the instantaneous logarithmic loss incurred by the forecaster $f_{i,t-1}$ at time $t-1$. Then, we write: $\mathbb{P}(y_t|f_{i,t-1}, D_{t-2}) = \exp\left(-l_{i,t-1}\right)$

Moreover, in order to build the index $\vartheta_{k,t}$ of arm k at time t, we propose three alternative definitions for the indices:

Sampling the expert distribution Like in Mellor and Shapiro (2013), the construction of the index $\vartheta_{k,t}$ is done according to a two-stage sampling process. First, one sample the discrete expert distribution $w_{i,t}$. Then, given the expert sampled $f_{i^\dagger,t}$, a sampling from the distribution $\mathbb{P}\left(\theta_t|D_{t-1}, f_{i^\dagger,t}\right)$ characterizes each arm k with a scalar $\theta_{k,i^\dagger,t}$. Finally, the index of each arm k is build according to the corrupt setting such that: $\vartheta_{k,t} = g_k^{-1}\left(\theta_{k,i^\dagger,t}\right)$

Bayesian Aggregation of experts Like in Alami et al. (2017), one can interpret eq (2) as a Bayesian aggregation of a growing number of experts seen as learning.
Thus, at each time step t, instead of having only one characterization for each arm k, a set $\Theta_{k,t} = \{\theta_{k,i,t} : \theta_{k,i,t} \sim \text{Beta}\left(\alpha_{k,i,t}, \beta_{k,i,t}\right)\ \forall\ i \in [\![1, t]\!]\}$ of t characterizations is available. Each element of $\Theta_{k,t}$ is a sampling from the Beta distribution associated to the model launched at time i. Finally, combining the Bayesian aggregation and the corrupt model of the environment leads us to build the index of each arm k as follows: $\vartheta_{k,t} = \frac{\sum_{i=1}^{t} g_k^{-1}(\theta_{k,i,t})w_{i,t}}{\sum_{i=1}^{t} w_{i,t}}$

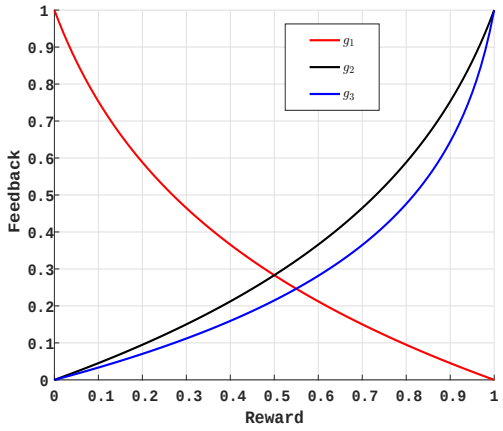
Picking the best estimated expert The easiest way to deal with a growing number of experts is to take at each time step the "best" expert in term of weight. Indeed, the Bayesian online change point detection tends to give to the expert starting at the last change point the highest weight. Thus, by letting $i^\star = \arg\max_i w_{i,t}$ be the best estimated expert at time t, the index of each arm k is built as follows: $\vartheta_{k,t} = g_k^{-1}\left(\theta_{k,i^\star,t}\right)$

Finally, by plugging one of the previous way to build the arm index $\vartheta_{k,t}$ into the formalism of Mellor and Shapiro (2013), we get the Global Switching Thompson Sampling with

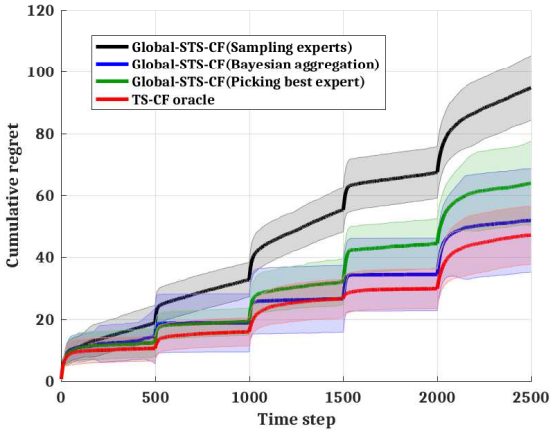
Corrupted Feedback (Global-STS-CF) described in algorithm 1. It should be noted that Global-STS-CF presents three variants according to the way of computing the arm index (Sampling the expert distribution, Bayesian aggregation of experts and picking the best estimated expert).

4. Experiments

In all the experiments, we consider a GS-CMAB of three arms observing five change points occurring at each 500 rounds. We compare the three versions of the Global-STS-CF with the TS-CF Oracle. Experiments are run 60 times.



(a) Mean-corruption functions ($g_{k,t}(\mu_{k,t}) = \nu_{k,t}$)



(b) Comparison between the three versions of Global-STS-CF and the oracle

Figure 1: Overall comparison of Global-STS-CF and the oracle.

First, the performances of the three variants of Global-STS-CF are very close to those of the oracle. This means that Global-STS-CF is able to perfectly deal with the switching environment. Then, using the Bayesian aggregation allows us to make performances challenging those of the oracle.

Discussion Observing figure 1(b), one should notice that Global-STS-CF is able to perfectly restart a TS-CF at each change point. This behavior is possible thanks to the inference model of the experts presented in eq (5). In fact, when a switch occurs the instantaneous gain $\mathbb{P}(y_t|f_{i,t-1}, D_{t-2})$ of all experts starting before the change point suddenly fall down because of their wrong estimation of the environment, giving the advantage to the experts newly created while annihilating the former ones. Then, the total mass of the expert distribution $w_{i,t}$ tends to focus around the *optimal* expert i.e. the expert starting at the most recent change point τ_κ and corresponds to the most appropriate characterization of the environment. This gives us the impression that Global-STS-CF restarts a new TS-CF at each change point.

5. Conclusion and future works

We have proposed Global-STS-CF: an extension with three variants of the Thompson Sampling for the Switching Corrupt Bandit Problem. From the experiments, the proposed algorithm presents excellent performances. It is worth noting that Global-STS-CF challenges the Thompson sampling with Corrupted Feedback oracle, an oracle which already knows the change points. These results arise from the fact that Global-STS-CF is based on the Bayesian concept of tracking the best experts which allows us to catch efficiently the change points. The proposed algorithm can naturally be extended to the *Per-arm Switching Corrupted Multi-Armed Bandit* by maintaining an expert distribution per arm. The next step of this work is to analyze the Global-STS-CF in term of pseudo cumulative regret.

Algorithm 1 Global **S**witching **T**hompson **S**ampling with **C**orrupted **F**eedback

procedure GLOBAL-STS-CF($\mathcal{K}, \mathcal{G}, T, \alpha_0, \beta_0, \rho$)
  $t \leftarrow 1, w_{1,t} \leftarrow 1$, and $\forall k \in \mathcal{K}\ \alpha_{k,1,t} \leftarrow \alpha_0,\ \beta_{k,1,t} \leftarrow \beta_0$ ▷ Initializations
  for $t \leq T$ **do** ▷ Interaction with environment
    $k_t \leftarrow$ CHOOSEARM $(\mathcal{G}, \{w\}_t, \{\alpha\}_t, \{\beta\}_t)$
    $y_t \leftarrow$ OBSERVEFEEDBACK(k_t) ▷ Bernoulli trial of parameter $g_{k_t}(\mu_{k_t,t})$
    $\{w\}_{t+1} \leftarrow$ UPDATEEXPERTMODEL $(\{w\}_t, \{\alpha\}_{k_t,t}, \{\beta\}_{k_t,t}, y_t, \rho)$
    $\{\alpha\}_{t+1}, \{\beta\}_{t+1} \leftarrow$ UPDATEARMMODEL$(\{\alpha\}_t, \{\beta\}_t, y_t, k_t)$
  end for
end procedure

---

procedure CHOOSEARM$(\mathcal{G}, \{w\}_t, \{\alpha\}_t, \{\beta\}_t)$
  $\forall\, k \in \mathcal{K}\ \forall\, i \in [\![1,t]\!]\ \ \theta_{k,i,t} \leftarrow \text{Beta}(\alpha_{k,i,t}, \beta_{k,i,t})$
  $\forall\, k \in \mathcal{K}\ \vartheta_{k,t} \leftarrow \begin{cases} g_k^{-1}(\theta_{k,i^\dagger,t}) & \text{Sampling the experts distribution} \\ \sum_{i\in[\![1,t]\!]} \frac{w_{i,t}}{\sum_{j\in[\![1,t]\!]} w_{j,t}}\ g_k^{-1}(\theta_{k,i,t}) & \text{Bayesian aggregation} \\ g_k^{-1}(\theta_{k,i^\star,t}) & \text{Picking the best expert} \end{cases}$
  return $\arg\max_k \vartheta_{k,t}$
end procedure
procedure UPDATEEXPERTMODEL$(\{w\}_t, \{\alpha\}_{k_t,t}, \{\beta\}_{k_t,t}, y_t, \rho)$
  $l_{i,t} \leftarrow -y_t \log\left(\frac{\alpha_{k_t,i,t}}{\alpha_{k_t,i,t}+\beta_{k_t,i,t}}\right) - (1-y_t)\log\left(\frac{\beta_{k_t,i,t}}{\alpha_{k_t,i,t}+\beta_{k_t,i,t}}\right) \forall\, i \in [\![1,t]\!]$
  $w_{i,t+1} \leftarrow (1-\rho)\, w_{i,t} \exp(-l_{i,t}) \forall\, i \in [\![1,t]\!]$ ▷ Increasing the size of expert $f_{i,t}$
  $w_{t+1,t+1} \leftarrow \rho \sum_i w_{i,t} \exp(-l_{i,t})$ ▷ Creating new expert starting at $t+1$
  return $\{w\}_{t+1}$
end procedure
procedure UPDATEARMMODEL$(\{\alpha\}_t, \{\beta\}_t, y_t, k_t)$
  $\alpha_{k_t,i,t+1} \leftarrow \alpha_{k_t,i,t} + \mathbb{1}(y_t = 1)\ \forall\, i \in [\![1,t]\!]$
  $\beta_{k_t,i,t+1} \leftarrow \beta_{k_t,i,t} + \mathbb{1}(y_t = 0)\ \forall\, i \in [\![1,t]\!]$
  $\alpha_{k,t+1,t+1} \leftarrow \alpha_0,\ \ \beta_{k,t+1,t+1} \leftarrow \beta_0\ \ \forall\, k \in \mathcal{K}$ ▷ Initializing new expert
  return $\{\alpha\}_{t+1}, \{\beta\}_{t+1}$
end procedure

References

Ryan Prescott Adams and David JC MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

Réda Alami, Odalric Maillard, and Raphael Féraud. Memory Bandits: a Bayesian approach for the Switching Bandit Problem. In *Neural Information Processing Systems: Bayesian Optimization Workshop.* , Long Beach, United States, 2017. URL `https://hal.archives-ouvertes.fr/hal-01811697`.

Pratik Gajane, Tanguy Urvoy, and Emilie Kaufmann. Corrupt bandits for preserving local privacy. In *ALT 2018-Algorithmic Learning Theory*, 2018.

Joseph Mellor and Jonathan Shapiro. Thompson sampling in switching environments with bayesian online change point detection. *CoRR, abs/1302.3721*, 2013.