Flexible, Efficient and Robust online experimentation platform using Multi-armed bandits

Jinjin Tian
Evidently Team, Amazon
Department of Statistics and Data Science, CMU
jinjint@amazon.com

Lenon Minorics
Causality Team, Amazon
minorics@amazon.com

Abstract

Online experimentation assumes evolving data stream, optimizing over metric or making inference at each time given the data collected so far. More and more experiments nowadays are running in this online fashion, either due to the restriction of data collecting process, or desire for flexibility. Particularly, for these online experiments, one would hope to be able to make decisions with confidence as soon as possible, without costing much revenue due to exploring bad treatments, while being sensitive to time variation in the treatment effect. Based on these motivations, we build an online experimentation platform, which is flexible for conducting different tasks (e.g. optimization or various testing tasks), efficient for reaching significance fast with low revenue cost, robust for handling time variation.

1 Introduction

Nowadays more and more companies runs experiments in a sequential manner due to the nature of high volume of data stream. Multi-armed bandits is a useful model for doing such online experiments: treating each arm as a treatment/variation, and adaptively pulling them to learn and exploit at the same time [1]. Running such experiments, people commonly are interested in either optimization or testing, which are in fact at odds: sampling strategy that is optimal for testing would hurt optimization, and vice versa. Different sampling strategies are carefully developed to achieve the optimal zone for either optimization or testing, where two main guidelines for optimization is Probability matching (e.g. Thompson Sampling) and Optimize under uncertainty (i.e. Upper Confidence Bound (UCB)); and Top-Two Thompson Sampling [1] and lil'UCB[2] for testing. However, the interpolation between those two are relatively under-explored. Degenne et al. [3] looked into this direction, by treating the uncertainty inflation level in lil'UCB[2] as the interpolation parameter, and studied how stopping time for best arm identification and accumulated regret will change with it. How the interpolation would work under many other testing problem like positive arm identification over control [4] using other sampling philosophy like Thompson Sampling remains under explored. Additionally, the environment in reality is constantly changing, causing an issue called time variation. This issue has been widely addressed for optimization goals, using ideas ranging from assuming an underlying stochastic process model [5] to change detection [6] and memory limitation [7]. However, how to do valid testing in this setting lacks formal investigation.

In this project, we first investigate the behavior of interpolation between optimization and testing using re-sampling sub-optimal or uncertainty inflation idea with Thompson sampling. From simulation, we find that, Thompson Sampling with uncertainty inflation is much more promising as a flexible adaptive sampling algorithm for both optimization and testing. For the abruptly time variation case, we propose a new change detector (CD) based on the any-time valid confidence interval, which

[1] Therefore, we will use the term *treatment* and *arm* exchangeably throughout the paper

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

is much faster and memory efficient compared with other current state of art change detectors, or memory decay methods. Using the policy of restarting if a change is detected, we are able to get much lower regret in either simulation or real dataset, together with much tighter any-time valid confidence bounds. In the end, use our findings, we are able to construct an flexible, efficient and robust online experimentation platform, using Thompson Sampling as adaptive sampling strategy for efficiency, with uncertainty inflation as convenient interpolation between optimization goal and testing goal for flexibility , and a novel change detector to protect against abrupt changes in treatment effect for robustness.

2 Problem set up: online experimentation

We assume users arrive sequentially, and at each time step t, we assign the arrived user to a treatment I_t from a set of possible treatments using only information up to time t, and record reward $Y_{I_t}(t)$. Throughout this paper, we consider stochastic reward, that is at each time t, $Y_i(t)$ associate with each arm $i \in [K]$ is a random variable follows some CDF $F_{i,t}$ with mean $\mu_i(t)$. Particularly, we call the environment stationary if $F_{i,t}$ does not changes with t and drops the index t if so; and non-stationary otherwise. Additionally, we assume users are independent from each other, that is $Y_i(1), Y_i(2), \dots,$ are independent random variables for each arm i. We denote $T_i(t)$ as the number of samples, and $\widehat{\mu}_{i,T_i(t)}$ as the empirical mean of rewards for treatment i by time step t. This online experimentation setting is most applicable to the system with large traffic and fast response. Large companies nowadays like Amazon, Google, Facebook fits this profile, and more and more are joining the club due to fast development of internet, transmission and automation. In the following, we formally introduce the two tasks of common interests in online experimentation:

Optimization The response Y is usually related to the commercial effect that people are interested in. For example, it can be the indicator of whether the user clicks on a given option or not on an webpage; the amount of money user spends on a product; the length of time a user stays with the recommended service. Optimizing of the reward is often transformed to minimization of the regret, which is the loss compared to always pull the optimal treatment: $R(T) = \sum_{t=1}^{T} \mathbb{E}\left[Y_{I_t^\star}(t) - Y_{I_t}(t)\right]$, where $I_t^\star$ is the optimal treatment at time t (i.e. the treatment with the highest mean at time t).

Inference Inference is essential when people would like to make decision with confidence. Most of time, promoting new changes or policies requires lots of time and effort, and aggressive decision can lead to countless loss. Here for simplicity and validness we consider inference within a stationary stage, that is the distribution of rewards does not change with time. Denote the baseline as μ_0, that is the mean of reward associated with the current status without any change, and confidence level δ, people are interested in (1) **Identify positive arms:** Discover a set of indices $\mathcal{S}_t$ at time t such that its false discovery rate (FDR) at time are controlled by δ, where FDR$(\mathcal{S}_t) = \mathbb{E}\left[\frac{|\mathcal{S}_t \cap \mathcal{H}_0|}{|\mathcal{S}_t| \vee 1}\right], \mathcal{H}_0 = \{i \in [K], \mu_i - \mu_0 \leq \epsilon\}$. (2) **Identify the best** k **arms:** Discover a set of indices $\mathcal{S}$ of size k, such that with probability $1 - \delta$, it is the set of the treatments with the top k means at time t, that is $\min_{i \in \mathcal{S}} \mu_i - \max_{j \in \mathcal{S}^c} \mu_j \leq \epsilon$.

3 Methods

The high level visualization of our proposed experimentation platform are shown in Section E, and detailed online experimetation algorithm is summarized in Algorithm 1, which allows flexible and efficient interpolation between optimization goal and various testing tasks listed in Section 2, and also able to protect against abrupt changes over time. Each solution/strategy we adopted in this platform are winners in our careful investigation and evaluation among many state of arts. In the following, we elaborates them one by one and reasoning about our choices.

3.1 Any-time valid inference and adaptive sampling

To do inference in the online setting, the normal way is via constructing any-time valid confidence interval for the true treatment effect, that is finding a sequence of $\{\mathcal{C}_i(T_i(t), \delta_i)\}_{t=1}^{\infty}$ for each treatment i with $\mathcal{C}_i(T_i(t), \delta_i) := [L_i(T_i(t), \delta_i), U_i(T_i(t), \delta_i)]$such that

$$\Pr\{\forall t \geq 1, L_i(T_i(t), \delta) \leq \mu_i \leq U_i(T_i(t), \delta_i)\} \geq 1 - \delta_i, \tag{1}$$

where $1-\delta_i$ is the confidence level assigned to treatment i. By definition, these any-time valid confidence intervals automatically corrects for the multiplicity issue introduced by data accumulation, allows for peeking at any time. Use the carefull design of δ_i in [8] and [4], we are able to respectively do testing like the best arm and positive arm identification with confidence $1-\delta$.

There are many ways to construct such any-time valid confidence interval, we particularly use the ones in [9] due to its proved optimality. Specifically, assuming $Y_{i,t} \overset{iid}{\sim} \text{subGaussian}(\sigma_i^2)$, then we have $[\widehat{\mu}_{i,T_i(t)} - \phi_i(T_i(t),\delta_i), \widehat{\mu}_{i,T_i(t)} + \phi_i(T_i(t),\delta_i)]$ satisfies (1), where

$$\phi_i(t,\delta) = \sqrt{\frac{\sigma_i^2}{t}\left[\log(\frac{1}{\delta}) + 3\log(\log(\frac{1}{\delta})) + 1.5\log\left(\log(et/2)\right)\right]}. \tag{2}$$

What comes with this any-time valid inference is the cost of time to reach significance, together with the loss of revenue. Adaptive sampling like MAB algorithms, on the other hand, can mitigate the time cost, and even revenue loss if design properly. In our paper, we use Thompson Sampling (TS) as the basic adaptive sampling algorithms, due to its superiority over other MAB algorithms in terms of optimization [10] (i.e. mininize revenue loss), and its compatibility for prior information. We start from TS and adjust it to accommodat different tasks (either purely optimization or inference) and scenarios.

3.2 Interpolating optimization and inference in stationary environment

The adaptive sampling strategy for optimization and inference goal turns out to be at odds, since optimization requires more exploitation while inference requires more exploration. In the following, we introduce two simple ideas of interpolating them. The first one is resampling the sub-optimal, that is at each time step t, with probability β_t, we sample the sub-optimal arm, and with probability $1-\beta_t$, we sample the optimal arm. Russo [1] used this idea to develop optimal strategy for best arm identification based on Thompson Sampling, and found that as $\beta_t \equiv \frac{1}{2}$, Thompson Sampling paired with resampling is optimal. Another idea is uncertainty inflation, that is inflate the uncertainty level for each arm by a certain amount α_t. [2] uses this idea to design optimal strategy for best arm identification based on UCB, and proved optimality up some constant. The common nice feature of these two ideas is that there are a simple tuning parameter controlling the level of interpolation. Section 4 shows the superioty of uncertainty inflation over re-sampling, therefore we adopt uncertainty inflation in our final platform.

3.3 Detect changes in peice-wise stationary environment

In order to do valid inference, we assume piece-wise stationary environment, that is the system changes only at a number of unknown time step, and remains stationary otherwise. There exists a class of methods trying to mitigate the influence of old past data gradually, e.g. parameter decay method in [7] or discounted-UCB, slide-window UCB in [11]. These methods are able to pick up the optimal arms correctly, but suffer from non-informative inference since they only allow limited memory of samples.

Another line of work that would allow better inference is based on change detection and restart, where the key is constructing powerful and efficient change detector. Among the current state-of-art change detectors, M-UCB [12] is not wanted in reality in the sense that it enforces a portion of uniform sampling, while ADSwitch [6] is very computationally expensive: it requires $O(t^2)$ computation at each time step t. GLR-klUCB [13] tried to use the generalized ratio test for change detection, but it still requires $O(t)$ computation at each time step t. In the following, we introduce a novel, simple, powerful change detector that only requires $O(1)$ computation at each time step, and do not require any portion of uniform sampling.

Change Detector (CD) Given a window size W, and $\phi_i(\cdot,\cdot)$ defined in (2), assume rewards follow subGaussian distribution, we decide there is a change occurred up to time t for a treatment i with confidence $1-\delta$ if

$$\left|\widehat{\mu}_{i,T_i(t)} - \widehat{\mu}_{i,T_i(t)-W}\right| > \frac{W}{T_i(t)}\left[\phi_i(T_i(t)-W,\delta/2) + \phi_i(W,\delta/2)\right], \tag{3}$$

where $\phi_i(\cdot,\cdot)$ is defined in (2). The selection of W can be decided if we assume minimal expected change in the mean Δ and variance σ_i^2: $W = \sigma_i^2/\Delta^2$.

One may refer to Section D for detailed reasoning. In Section 4.2, we showed that our change detector has much lower regret compared with parameter decay strategy, and is able to restart almost right after the change point; we also use a real dataset to show that our change detector has much lower regret compared with other solutions to the problem, and gives near optimal inference of the best arm.

4 Experiments

The following results using the independent Beta-Bernoulli model, and the extra experiments for independent Normal-Normal model are in Section B for interested readers. Specifically, in this section we assume each arms are independent from each other, and $Y_i(t) \sim \text{Bernoulli}(\mu_i(t))$ across all t for each arm i, where we assume $\mu_i(t)$ is fixed numbers in $\mathbb{R}$ and is piece-wise linear in time t. In the sampling step, we using the Beta prior for Thompson Sampling, that is $\theta_i^t \sim \text{Beta}(a_i^t, b_i^t)$ such that the posterior can be easily updated using conjugate relationship. Particularly, for $\pi_i^t = \text{Beta}(a_i^t, b_i^t)$, the posterior given observation $Y_i(t)$ can be updated using $\pi_i^{t+1} = \text{Beta}(a_i^t + Y_i(t), b_i^t + 1 - Y_i(t))$. One may refer to Section A for details of conducting resampling and inflation in this simple model. The following results are all averaged over 30 trials. Each setting is ran under the confidence level $\delta = 0.05$, and precision $\epsilon = 0$.

4.1 Evaluating the interpolation effect in stationary case

In this section, we would like to evaluate the effect of interpolation using either resampling suboptimal or inflating posterior variance in the stationary environment. We run the algorithms with some maximum pulls, and record the time when reached statistical significance as the stopping time. If one algorithm have not reach significance until the maximum pulls, the stopping time will be NA. We compare the performance of Thompson sampling with different interpolation parameter ($\beta \in \{0.5, 0.4, 0.3, 0.2, 0.1, 0\}$ for resampling suboptimal, $\alpha \in \{10, 5, 4, 3, 2, 1\}$ for inflate posterior variance), when dealing with different inference problems. We use uniform sampling as a benchmark.

Best arm identification. We first look at a problem of identifying the best arm among five arms with mean $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. In Figure 5, we plot the boxplot of stopping time and regret at stop[2] versus different interpolation parameter, with uniform sampling as a benchmark in the first position, using results of 30 independent trials. We can see that from Figure 5 that, resampling suboptimal strategy underperforms uniform sampling with respect of either inference or optimization (i.e. either stopping time or regret at stopping), while inflating variance strategy outperforms uniform sampling in both aspect. Figure 5 also demonstrate interesting influence of the interpolation parameter, when under none interpolation (that is 0 for resmapling and 1 for inflating), the stopping time will be infinite; however just by a bit from the end of none interpolation, the algorithm will have finite stopping time; the reduction in stopping time seems to be logarithmic in the interpolation parameter, that is setting the interpolation parameter too far from none will not bring much reduction in stopping time, while it will hurt regret on the other hand.

Positive arm identification. Then we look into the testing problem of identifying positive arms with FDR control. We consider 20 arms with means randomly generated from $U[0, 1]$, and try to identify arms that is better the control level 0.5. We ran the algorithms with maximum $5 * 10^5$ pulls, and plot the power[3] versus time (i.e. the number of pulls) in Figure 6, with different color marking different interpolation level. We can see that, given the range of interpolation parameter we tried, both resampling optimal or inflating the variance strategy underperforms uniform sampling in terms of power, while outperforms in terms of regret. However, inflating the variance looks more promising than resampling once again, since its power at level $\alpha = 10$ is able to fast converge to that of uniform sampling, while its regret is much lower than either resampling suboptimal or uniform sampling.

In conclusion, inflating the posterior variance serves as a better interpolation strategy compared with resampling the sub-optimal, from the aspect of reducing time to reach significance without cost of much regret in both inference of best arm or positive arm identification. We provide more evidence for supporting this argument in Section C from the aspect of pulling probability of different arms. On the other hand, Thompson sampling performs similarly as UCB in term of serving as foundations for doing interpolation, while Thompson sampling shows slightly stronger trade-off

[2] If the stopping time is NA, then the regret at stop will be recorded as the regret at the maximum pulls.

[3] the proportion of signals that are correctly identified with confidence

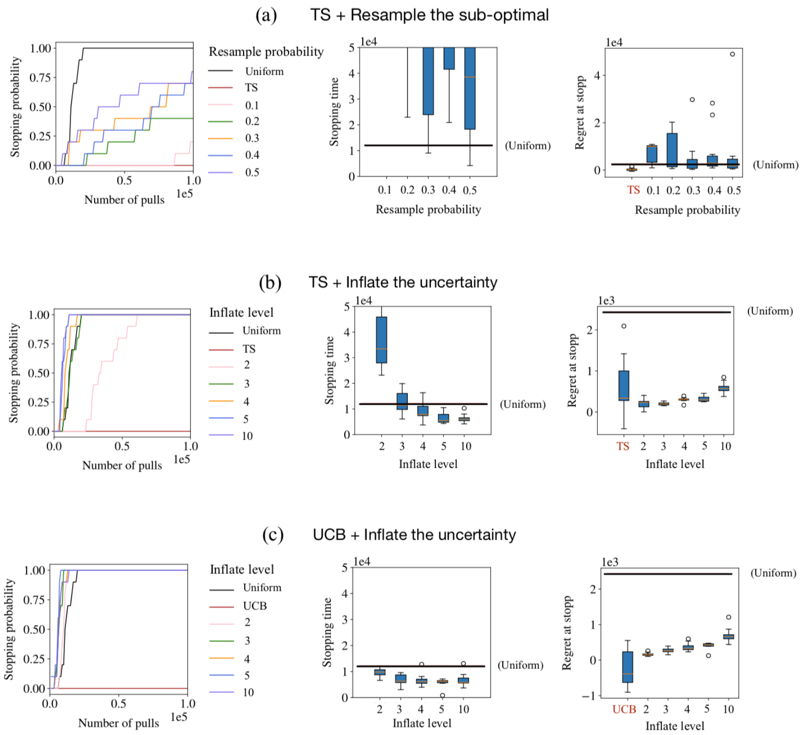


Figure 1: The accumulated stopping probability (proportion of experiments that have stopped out of total trials) and boxplot of stopping time (time when reaches statistical significance) and regret at stop (empirical gap of accumulated reward from always pulling the best) versus different interpolation parameter, when finding the best arm among five arms with means $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. Particularly, (a) plots for Thompson Sampling with resampling probability $\beta \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ and uniform sampling on the left in each boxplot, (b) plots for the Thompson Sampling with inflated variance by different level $\alpha \in \{1, 2, 3, 4, 5, 10\}$, and (c) plots for the UCB with inflated variance by different level $\alpha \in \{1, 2, 3, 4, 5, 10\}$.

effect (i.e. same level of uncertainty inflation leads to more reduction of stopping time when using Thompson Sampling). Given many other known practical advantages of Thompson Sampling like it allows of easy incorporation with prior information and dependency structure among arms, we recommend going with Thompson Sampling in practice. On a higher level, as for comparison between adaptive sampling and uniform sampling generally, except for the obvious advantage of adaptive sampling with regard revenue cost, it sometimes also outperforms uniform sampling in terms of time to reach significance. For example, when dealing with best arm identification, or trying to achieve moderate power in inferring behaviour of many arms (i.e. positive arms identification). However, adaptive sampling will leads to longer time comparing with uniform sampling to reach high power in inferring behaviour of many arms, which is as expected. Therefore we suggest one should put the types of tasks and scenarios into serious consideration when choosing to use adaptive sampling carefully in practice.

4.2 Evaluating change detector in peice-wise stationary case

In this section, we would like to evaluate the performance of our change detector proposed in Section 3.3. We consider four sampling strategies related to Thompson Sampling (TS): TS, TS with restarts as if the change point is known (TS-Oracle), TS with the change detector proposed in Section 3.3 (TS-CD), TS with parameter decay [7] (TS-Decay); and additionally the states-of-art methods reviewed in Section 3.3, together with other relevant existing work [11, 14]. We again assume Beta-Bernoulli model as in Section 4.1, but we consider that the means of arms change abruptly at certain unknown time points, while remains unchanged between each change time points.

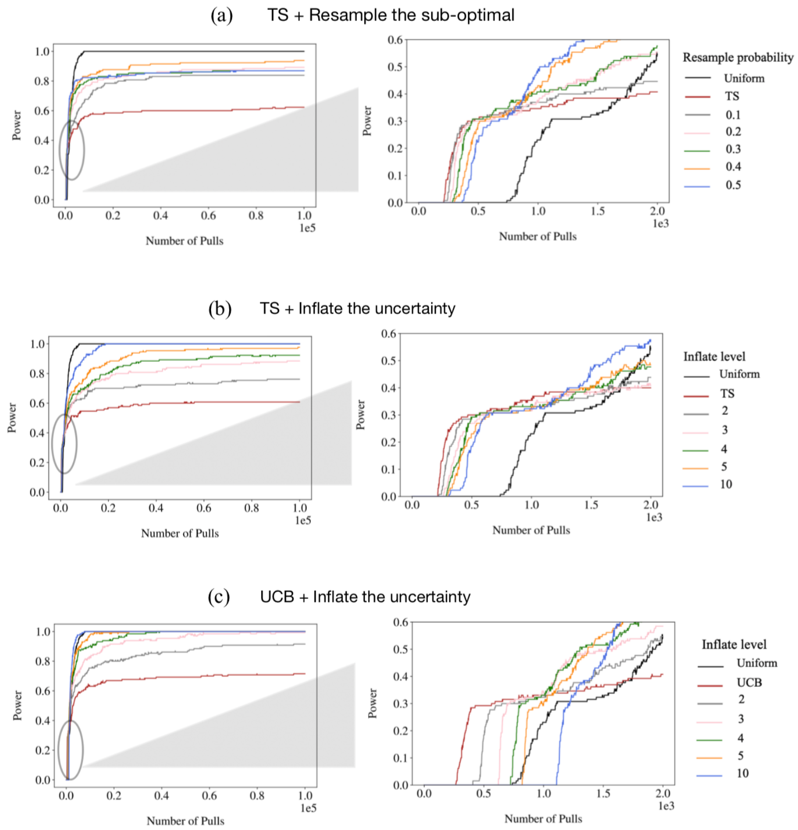


Figure 2: The power versus time given different interpolation parameter, when finding the arms better than the control level 0.5 with FDR control under 0.05 among 20 arms with means randomly generated from $U[0, 1]$. Particularly, (a) plots for Thompson Sampling with resampling probability $\beta \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$, (b) plots for the Thompson Sampling with inflated variance by different level $\alpha \in \{1, 2, 3, 4, 5, 10\}$, and (c) plots for the Thompson Sampling with inflated variance by different level $\alpha \in \{1, 2, 3, 4, 5, 10\}$.

For the sake of simplicity, we do not include the effect of interpolation, and only evaluate the performance by regret versus time, and the accurateness of inference of the best arm (whose value and identity may changes across the time). Particularly, we use the any-time valid confidence bound to represent our inference on the best arm.

Experiment on synthetic data We first consider using purely synthetic data, where we generate 5 arms with means change with time as in Figure 3(a). Particularly, we use $W = 1000$ in TS-CD, and decay rate $\gamma = 0.001$ in TS-Decay, which are selected by optimality within a range of different options. We plot in Figure 3(c) the any-time valid confidence bounds for the true mean of best arm in each stationary period, together with the empirical estimation and the true value, and plot in Figure 3(b) the regret versus time for each sample strategy. One can see that, our TS-CD is able to detect changes almost right after the change points, thus giving the most accurate inference of the best arm comparing with others, together with regret much lower than others as well.

Experiment on Yahoo! Front Page Dataset Then we evaluate the performance of our TS-CD algorithm using the benchmark dataset publicly published by Yahoo! [4]. This dataset provides a binary value for each arrival to represent whether the user clicks the specified article. We follows we use one arm to represent one article and assume a Bernoulli reward, and obtain a piece-wise stationary scenario for six articles with $T = 4 \times 10^5$, and 8 change points, as shown in Figure 4(a), following the standard processing steps in literature of abrupt change time variation [12]. Particularly, we use $W = 500$ in TS-CD, and decay parameter $\gamma = 0.01$ in TS-Decay. Again, we plot the regret

[4] Yahoo! Front Page Today Module User Click Log Dataset

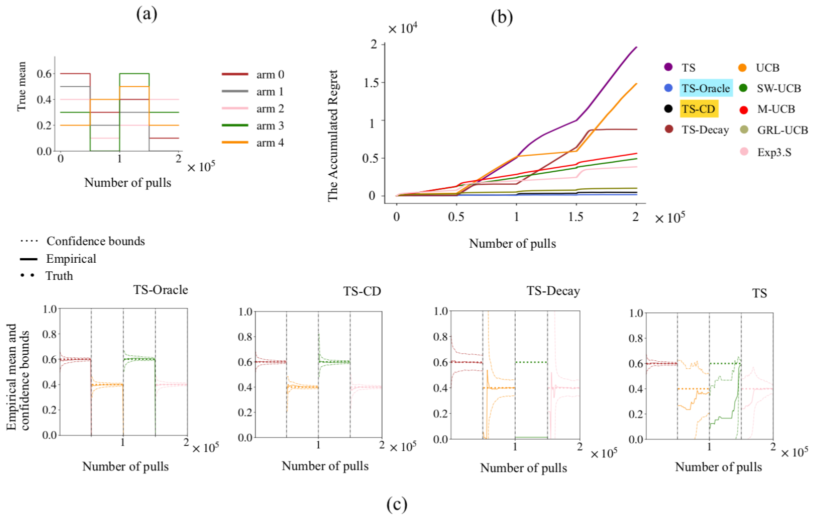


Figure 3: The performance of TS-CD when facing simulated abrupt time variation. (a) plots the true mean of arms versus time; (b) plots the regret versus time; (c) plots the inference of the best arm during each stationary period, with bold line indicating the empirical estimation, bold dashed line indicating the truth and dashed line indicating the confidence bounds.

versus time in Figure 3(b), which shows that our TS-CD strategy is close to the optimal solution TS-Oracle, while being much better than other state-of-arts methods. Figure 3(c) plots the inference of the best arm over each stationary period for TS-CD and TS-Oracle, which shows that our TS-CD gives confidence interval covers the true mean most of times, with empirical estimate close to the true mean if failed to cover. Overall, the inference of the best arm of TS-CD is arguably comparable to the optimal solution, TS-Oracle.

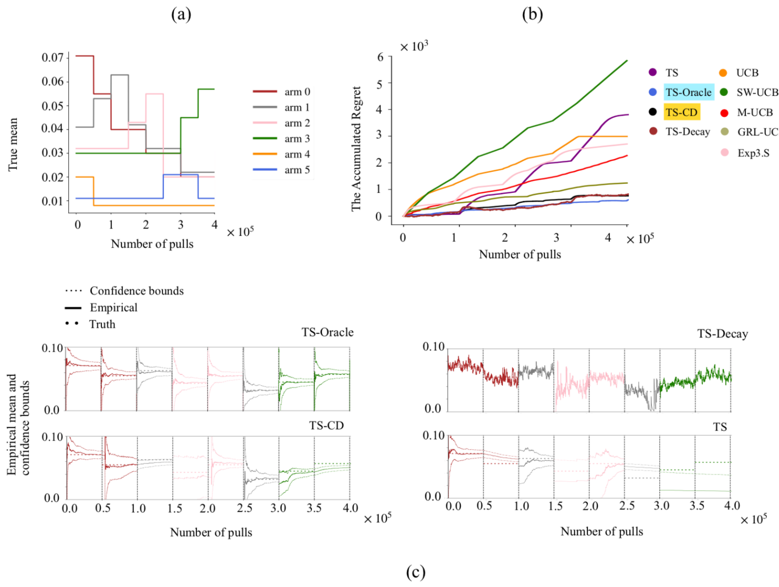


Figure 4: The performance of TS-CD when dealing with real world problem generated from Yahoo. (a) plots the true mean of arms versus time; (b) plots the regret versus time using different state-of-arts algorithms designed for handling time variation; (c) plots the inference of the best arm during each stationary period, with bold line indicating the empirical estimation, bold dashed line indicating the truth and dashed line indicating the confidence bounds.

5 Conclusion

In conclusion, from the above investigation, we find Thompson Sampling with uncertainty inflation a promising strategy to do interpolation between optimization and inference; and we propose a novel change detector to force the restart to deal with the abrupt changing non-stationary environment, which proves working well in either synthetic or real problem. We summarize our findings in to a general platform for flexible online experiments with multiple goals, and robustness for time variation. We hope this platform can help people to do efficient testing with low cost in various environment.

References

[1] Daniel Russo. Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418, 2016.

[2] Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439, 2014.

[3] Rémy Degenne, Thomas Nedelec, Clément Calauzènes, and Vianney Perchet. Bridging the gap between regret minimization and best arm identification, with application to a/b tests. *arXiv preprint arXiv:1810.04088*, 2018.

[4] Kevin Jamieson and Lalit Jain. A bandit approach to multiple testing with false discovery control. *arXiv preprint arXiv:1809.02235*, 2018.

[5] Djallel Bouneffouf and Raphael Féraud. Multi-armed bandit problem with known trend. *Neurocomputing*, 205:16–21, 2016.

[6] Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory*, pages 138–158, 2019.

[7] Vishnu Raj and Sheetal Kalyani. Taming non-stationary bandits: A bayesian approach. *arXiv preprint arXiv:1707.09727*, 2017.

[8] Fanny Yang, Aaditya Ramdas, Kevin G Jamieson, and Martin J Wainwright. A framework for multi-a (rmed)/b (andit) testing with online fdr control. In *Advances in Neural Information Processing Systems*, pages 5957–5966, 2017.

[9] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Uniform, nonparametric, non-asymptotic confidence sequences. *arXiv preprint arXiv:1810.08240*, 2018.

[10] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

[11] Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415*, 2008.

[12] Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 418–427, 2019.

[13] Lilian Besson and Emilie Kaufmann. The generalized likelihood ratio test meets klucb: an improved algorithm for piece-wise non-stationary bandits. *arXiv preprint arXiv:1902.01575*, 2019.

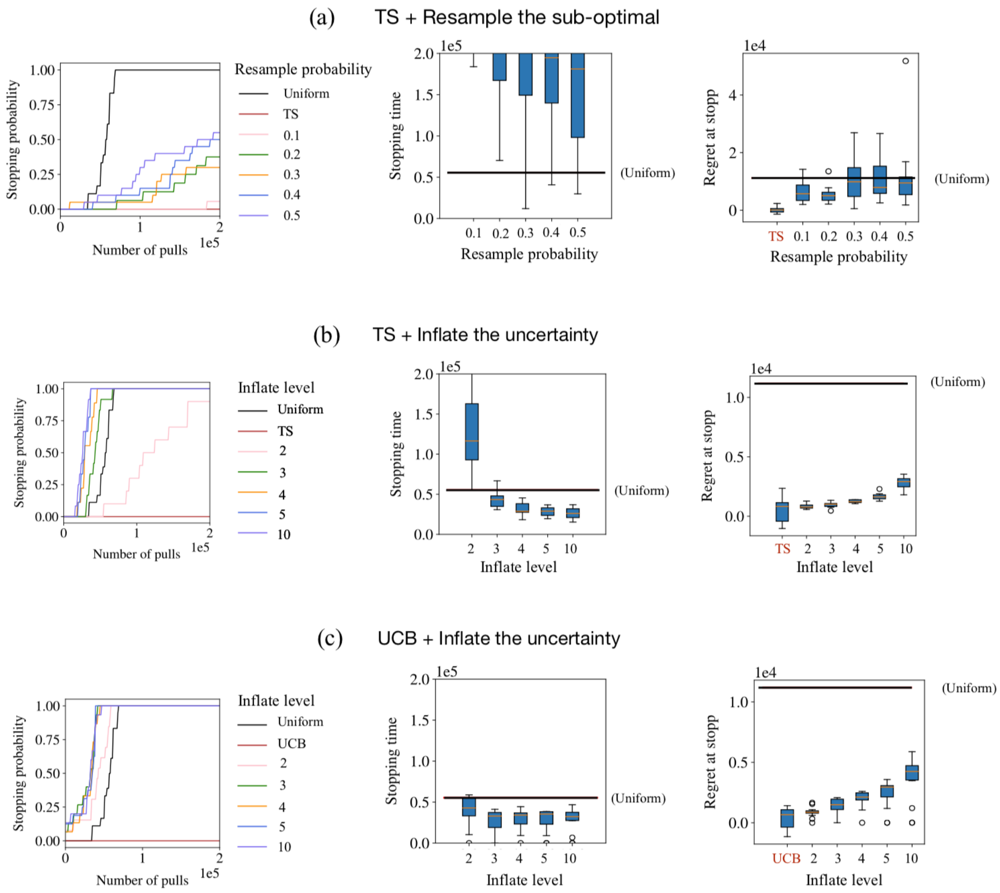
[14] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Optimal exploration–exploitation in a multi-armed bandit problem with non-stationary rewards. *Stochastic Systems*, 9(4):319–337, 2019.

Appendices

Appendix A Details about interpolation under Beta Bernoulli model

Under such simple model in Section 4.1, we incorporate with resampling suboptimal with probability $\beta \in [0, 0.5]$ by the following: at each time step t, after sampling each θ_i^t from the posterior distribution π_i^t, we pull best arm $i^\star$ (i.e. the one with highest θ_i^t) with probability $1 - \beta$, and we resample until the the optimal arm is different from $i^\star$. We we incorporate with uncertainty inflation with level $\alpha \in [1, \infty)$ as the following: at each time step, we sampling each θ_i^t from modified posterior distribution $\widetilde{\pi}_i^t = \text{Beta}(a_i^t/\alpha^2, b_i^t/\alpha^2)$, such that the posterior variance will be inflating as multiplying α^2 with posterior mean unchanged. Note that, resampling with probability $\beta = 0$ is equivalent to inflating variance by $\alpha = 1$ times, both recovers Thompson Sampling.

Appendix B Additional experiments for Gaussian arms



(a) TS + Resample the sub-optimal

(b) TS + Inflate the uncertainty

(c) UCB + Inflate the uncertainty

Figure 5: The accumulated stopping probability (proportion of experiments that have stopped out of total trials) and boxplot of stopping time (time when reaches statistical significance) and regret at stop (empirical gap of accumulated reward from always pulling the best) versus different interpolation parameter, when finding the best arm among five arms with means $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. Particularly, (a) plots for Thompson Sampling with resampling probability $\beta \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ and uniform sampling on the left in each boxplot, (b) plots for the Thompson Sampling with inflated variance by different level $\alpha \in \{1, 2, 3, 4, 5, 10\}$, and (c) plots for the UCB with inflated variance by different level $\alpha \in \{1, 2, 3, 4, 5, 10\}$.

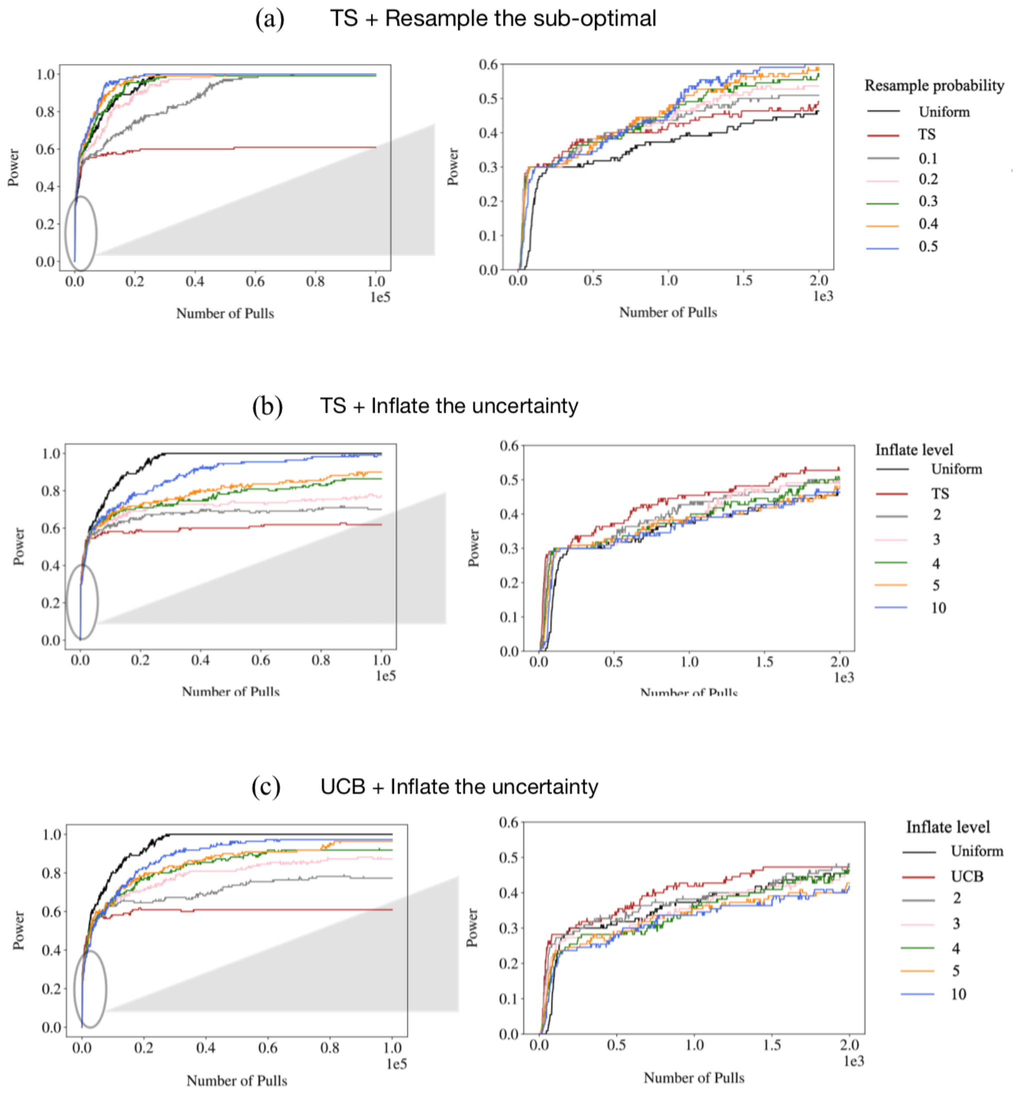


Figure 6: The power versus time given different interpolation parameter, when finding the arms better than the control level 0.5 with FDR control under 0.05 among 20 arms with means randomly generated from $U[0, 1]$. Particularly, (a) plots for Thompson Sampling with resampling probability $\beta \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$, (b) plots for the Thompson Sampling with inflated variance by different level $\alpha \in \{1, 2, 3, 4, 5, 10\}$, and (c) plots for the Thompson Sampling with inflated variance by different level $\alpha \in \{1, 2, 3, 4, 5, 10\}$.

In this section we assume each arms are independent from each other, and $Y_i(t) \sim \text{Normal}(\mu_i(t), 1)$ across all t for each arm i, where we assume $\mu_i(t)$ is fixed numbers in $\mathbb{R}$ and is piece-wise linear in time t. In the sampling step, we using the Normal prior for Thompson Sampling, that is $\theta_i^t \sim \text{Normal}(a_i^t, b_i^t)$ such that the posterior can be easily updated using conjugate relationship. Particularly, for $\pi_i^t = \text{Normal}(a_i^t, b_i^t)$, the posterior given observation $Y_i(t)$ can be updated using $\pi_i^{t+1} = \text{Normal}(\frac{1}{1+b_i^t} a_i^t + \frac{b_i^t}{1+b_i^t} Y_i(t),\ \frac{b_i^t}{1+b_i^t})$. The following results are all averaged over 30 trials. Each setting is ran under the confidence level $\delta = 0.05$, and precision $\epsilon = 0$.

Appendix C Reasoning the superiority of uncertainty inflation over resampling sub-optimal

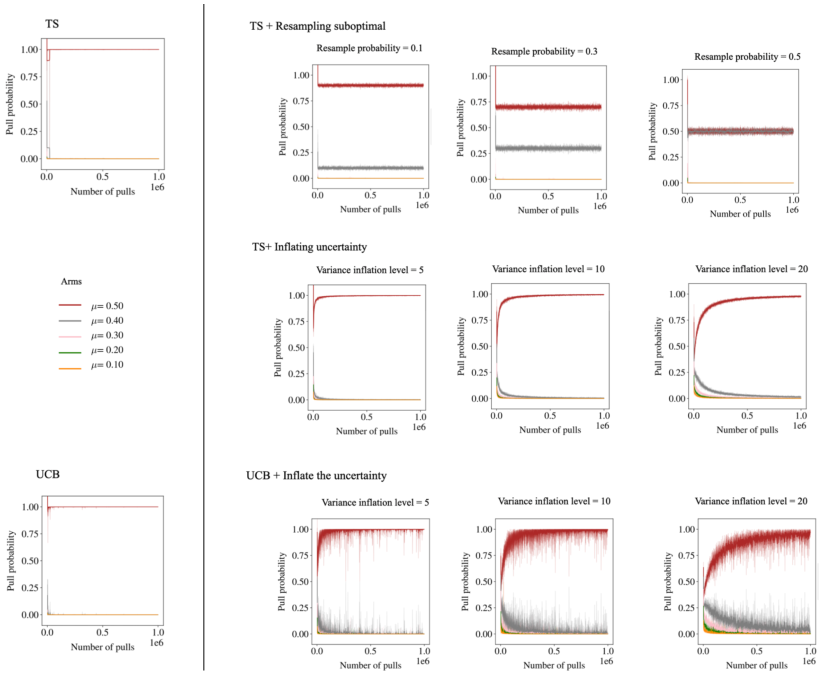


Figure 7: The empirical probability of pulling each arms for different sampling policy. One may see that resampling suboptimal results in the probability of pulling the best arm not converge to one, a behaviour that is unwanted in regret minimization.

Appendix D Reasoning of the Change Detector in Section 3.3

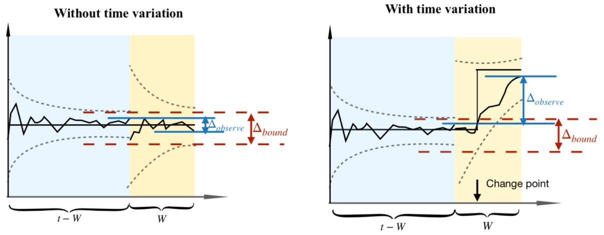


Figure 8: Intuition behind the proposed change detector in Section 3.3.

At each time step, one would like to know whether a change has occurred, we would test this using contradiction. Assume no change has occurred up to time t for arm i and that $Y_i(t)$ is sub-Gaussian random variable with the same mean μ. When $T_i(t) > W$, from triangle inequality, we have that

$$\left|\frac{1}{T_i(t)-W}\sum_{q=1}^{T_i(t)-W} Y_i(q) - \frac{1}{W}\sum_{s=1}^{W} Y_i(T_i(t)-W+s)\right|$$
$$\leq \left|\frac{1}{T_i(t)-W}\sum_{q=1}^{T_i(t)-W} Y_i(q) - \mu\right| + \left|\mu - \frac{1}{W}\sum_{s=1}^{W} Y_i(T_i(t)-W+s)\right| \tag{4}$$

From (2), we know that, with probability less than $2 * \delta/2 = \delta$, we have

$$\left|\frac{1}{T_i(t)-W}\sum_{q=1}^{T_i(t)-W} Y_iq - \mu\right| \geq \phi_i(T_i(t)-W,\delta/2), \quad \text{or} \quad \left|\mu - \frac{1}{W}\sum_{s=1}^{W} Y_i(T_i(t)-W+s)\right| \geq \phi_i(W,\delta/2), \tag{5}$$

which is saying with probability bigger than $1-\delta$, we have

$$\left|\frac{1}{T_i(t)-W}\sum_{q=1}^{T_i(t)-W} Y_i(q) - \mu\right| \leq \phi_i(T_i(t)-W,\delta/2), \quad \text{and} \left|\mu - \frac{1}{W}\sum_{s=1}^{W} Y_i(T_i(t)-W+s)\right| \leq \phi_i(W,\delta/2). \tag{6}$$

Therefore, with probability bigger than $1-\delta$, we have

$$\left|\frac{1}{T_i(t)-W}\sum_{q=1}^{T_i(t)-W} Y_i(q) - \frac{1}{W}\sum_{s=1}^{W} Y_i(T_i(t)-W+s)\right| \leq \phi_i(T_i(t)-W,\delta/2) + \phi_i(W,\delta/2), \tag{7}$$

since

$$\begin{aligned}\left|\frac{1}{T_i(t)-W}\sum_{q=1}^{T_i(t)-W} Y_i(q) - \frac{1}{W}\sum_{s=1}^{W} Y_i(T_i(t)-W+s)\right| &= \left|\widehat{\mu}_{i,T_i(t)-W} - \frac{W\widehat{\mu}_{i,T_i(t)} - (T_i(t)-W)\widehat{\mu}_{i,T_i(t)-W}}{W}\right| \\ &= \left|\frac{T_i(t)}{W}\widehat{\mu}_{i,T_i(t)} - \widehat{\mu}_{i,T_i(t)-W}\right|, \end{aligned} \tag{8}$$

we have (7) simplified to that, with probability bigger than $1-\delta$,

$$\left|\widehat{\mu}_{i,T_i(t)} - \widehat{\mu}_{i,T_i(t)-W}\right| \leq \frac{W}{T_i(t)}\left[\phi_i(T_i(t)-W,\delta/2) + \phi_i(W,\delta/2)\right]. \tag{9}$$

Appendix E Online experimentation flow

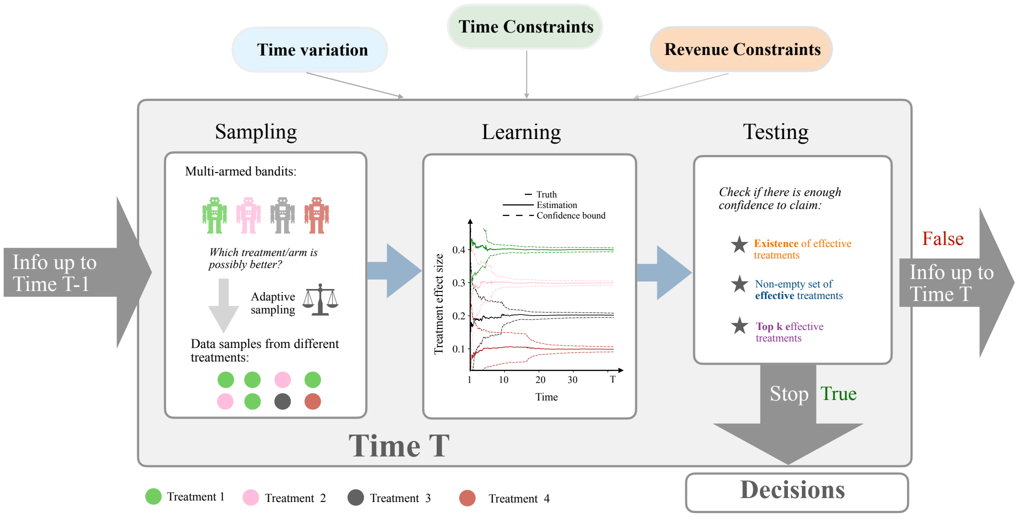


Figure 9: Concept map of the platform.

Appendix F Detailed online experimentation algorithm

Algorithm 1: The pseudo online experiment algorithm

```
Input: Initial prior $\pi^1$, number of treatments $K$, output size $m$, control level $\mu_0$, precision
       parameter $\epsilon$, confidence level $\delta$, variance inflation level $\alpha \in [1, \infty)$, time detection
       window size $W$, confidence corrected for dependency $\delta' = \frac{\delta}{6.4 \log 36/\delta}$, any-time valid
       confidence functions $\phi_i(\cdot, \cdot)$ for each arm.
for $t = 1, 2, \ldots$ do
    Sampling:
    Sample from variance inflated posterior: $(\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_K) \sim \pi^t(\alpha)$;
    Get the indexes with $m$ highest mean: $\mathcal{I}_t \leftarrow \arg\max_{\mathcal{I} \in m^{[K]}} \{\widehat{\theta}_i\}_{i \in \mathcal{I}}$;
    Pull treatments $\mathcal{I}_t$, get reward $\{Y_i(t)\}_{i \in \mathcal{I}_t}$.
    Check for abrupt change:
    for $i \in \mathcal{I}_t$ do
        if $\mod (T_i(t), W) = 0$ and $T_i(t) > W$ then
            if $|(\widehat{\mu}_{i,T_i(t)} - \widehat{\mu}_{i,T_i(t)-W})| > \frac{W}{T_i(t)} (\phi(T_i(t) - W, \delta/2m) + \phi(W, \delta/2m))$ then
                Restart;
            end
        end
    end
    Updating:
    Update posterior: $\pi^{t+1} \leftarrow \mathcal{F}(\pi^t, \{Y_i(t)\}_{i \in \mathcal{I}_t})$;
    Update empirical mean: $\widehat{\mu}_{i,T_i(t+1)} \leftarrow \frac{1}{T_i(t)+\mathbf{1}\{i=I_t\}} \left(T_i(t)\widehat{\mu}_{i,T_i(t)} + Y_t \mathbf{1}\{i = I_t\}\right)$ for all $i$;
    Update set of positive treatments:
      $s_t(k) := \{i \in [K], \widehat{\mu}_{i,T_i(t)} + \phi_i(T_i(t), \frac{\delta' k}{K}) \geq \mu_0\}$;
      $\mathcal{A}_t = \{k \in [K] : |s_t(k)| \geq k\}$;
      $\mathcal{S}_{t+1} = s_t(\max\{\mathcal{A}_t\})$.
    Update set of best $m$ treatments:
      $\mathrm{LCB}_i(t) := \widehat{\mu}_{i,T_i(t)} - \phi_i(T_i(t), \frac{\delta}{2(K-m)})$,        $\mathrm{UCB}_i(t) := \widehat{\mu}_{i,T_i(t)} + \phi_i(T_i(t), \frac{\delta}{2m})$
      $\mathcal{U}_t = \arg\max_{\mathcal{U} \in m^{[K]}} \{\widehat{\mu}_{i,T_i(t)}\}_{i \in \mathcal{U}}$,    $\mathcal{L}_t = \arg\max_{\mathcal{L} \in m^{[K]}, \mathcal{L} \neq \mathcal{U}_t} \{\mathrm{UCB}_i(t)\}_{i \in \mathcal{L}}$.
    Check for output:
    if $\min \mathrm{LCB}_{\mathcal{U}_t}(t) > \max \mathrm{UCB}_{\mathcal{L}_t}(t) - \epsilon$ then
        return $\mathcal{U}_t$ as the best $m$ treatments.
    end
    if $|\mathcal{S}_t| \geq m$ then
        return $\mathcal{S}_t$ as the positive $m$ treatments over control with FDR control.
    end
end
```