Hugo Mallet
hugo.mallet@polytechnique.edu
Mithuran Gajendran
mithuran.gajendran@polytechnique.edu

A framework for Multi-A(rmed)/B(andit)
testing with online FDR control

1 Introduction

The paper A framework for Multi-A(rmed)/B(andit)testing with online FDR control[4] proposes an alternative framework to existing setups for controlling false alarms when multiple A/B testings are run over time. This article aims to summarize and explain more intuitively the main idea introduced as well as questioning ourselves about potential limitations.

2 Context

A/B testing is a procedure that aims to measure the impact of a version change of a variable on the achievement of a goal. It works as follows :
Given a default configuration and several alternatives, the standard practice is to divert a small amount of scientist-traffic to a randomized trial over these alternatives and record the desired metric for each of them. If an alternative appears to be significantly better, it is implemented; otherwise, the default setting is maintained. However, this methods has some drawbacks that this paper aims to overcome :

Drawbacks of A/B testing :

- **Adaptive traffic** : While proposing alternatives, we can quickly infer that some are clearly less efficient than others and as stated previously, A/B(/n) testing frameworks allocates traffic uniformly over models. Hence, companies running A/B tests are susceptible to spend time and resources on bad alternatives whereas it would be smarter to confer more traffic to the best alternative.
In order to deal with this drawback, one can employ adaptive sampling algorithms like Multi-armed bandit.
- **Continuous monitoring** : Ongoing A/B testings' criterions cannot be modified and one have to wait till the end of the whole process to get feedbacks from it. However, it would be useful to be able to adjust termination criteria as time goes by and possibly stop earlier or later than originally intended for the A/B test. More precisely, the practice of continuous monitoring can easily fool the tester to believe that a result is statistically significant, when in reality it is not.
The introduction of notions such as adaptive null hypothesis and always valid p-value are used in this paper to deal with this drawback
- **Controlling the False Discovery Rate** : In A/B testing, the lack of sufficient evidence or an insignificant improvement of the metric may make it undesirable from a practical or financial perspective to replace the default alternative. Therefore, when a company runs hundreds to thousands of A/B tests within a year, ideally the number of statistically insignificant changes that it made should be small compared to the total number of changes made. Controlling the false alarm rate of each individual test at a desired level however does not achieve this type of control, also known as controlling the **False Discovery Rate**. It might be also desirable to detect better alternatives, and to do so as quickly as possible.
This concern can be handled using recent advances in online False Discovery Rate (FDR) control.

In order to deal with those drawbacks, this paper introduces a combined framework that can be described as **doubly-sequential** : sequences of Multi-armed bandit tests, each of which is itself sequential.

This ”Meta-framework” uses and combines the three particularity described earlier :

- Adaptive Sampling thanks to Multi-Armed Bandit algorithm
- Continuous Monitoring
- Online FDR procedures

3 Source of the idea

In order to fully understand this framework, we first need to understand the main notations and algorithms which are the stepping stone of all the results that this paper introduces.

p-values and null hypothesis :

In statistics, the p-value is the probability of obtaining the observed results of a test, assuming that the **null hypothesis** is correct. It is the level of marginal significance within a statistical hypothesis test representing the probability of the occurrence of a given event. The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis. We can set the significance point/level as α. Hypotheses with p-values below a significance level α (which is often set to 0.05) are considered to be **statistically significant**. For example, if the p-value is equal to 0.03, there would be 3 percent chance of obtaining our results knowing that the null hypothesis is true. This methods is widely used in real life use cases where one needs to prove the statistical significance of a scenario.

False Discovery Rate :

Each experiment can be viewed as a **test of a null hypothesis**. In our case the null hypothesis can be seen as the claim that the current control arm is the best. Hence, we call discovery the **claim that an alternative arm is the best** (rejected null hypothesis). However, and erroneous claim is called a **false discovery**.
Hence, we can introduce the **False Discovery Rate (FDR)** as the expected ratio of false discoveries to the total number of discoveries. In simple words, it is the rate that an arm that was claimed to be the best (significant hypothesis) was in fact erroneous.

$$FDR = \mathbb{E}[\frac{Q}{N}] \tag{1}$$

With :

- Q = number of False Discoveries
- N = total number of Discoveries

An alternative to the FDR is the **modified False Discovery Rate** (mFDR) which is the ratio of expected number of false discoveries to the expected number of total discoveries

$$mFDR = \frac{\mathbb{E}[Q]}{\mathbb{E}[N]} \tag{2}$$

Both those metrics are standard quantities for multiple testing applications, which is one of the constraint of our Framework. Furthermore, this paper aims to control the **False Discovery Rate** in an **online manner** by considering an ordered, sequential number of null hypothesis. This procedure aims to control **type I errors** (false positives) by using an adaptive significant α level for each hypothesis.

Multi-Armed Bandit :

In machine learning, the “exploration vs. exploitation tradeoff” applies to learning algorithms that try to acquire new knowledge and maximize their reward at the same time (which is part of Reinforcement Learning). In this setting, regret is defined as a decrease in reward due to executing the learning algorithm instead of behaving optimally from the very beginning. Algorithms that optimize for exploration tend to incur more regret. A/B testing is purely exploratory problem thus, A/B tests can have very high regret.

Multi-armed bandit algorithms can be thought as an improvement to A/B testing that balances exploitation and exploration during the learning process.
A Multi-Armed Bandit solution uses the results from its own exploration to allocate more traffic to variants that are performing well, while allocating less traffic to variants that are underperforming. There are several Multi-Armed Bandit algorithms, each favoring exploitation over exploration to different degrees. Three of the most popular are Epsilon Greedy, Thompson Sampling, and Upper Confidence Bound 1.

4 Important Results

This paper introduces two results that, we believe, are the most important :

- How to embed Multi-Armed Bandit algorithms into an online FDR framework
- Guarantees that this introduced framework is efficient.

1) Embed Multi-Armed Bandit algorithms into an online FDR framework

FDR online procedure consists in considering an ordered and possibly infinite sequence of null hypotheses H where, at each step i, the statistician must decide whether to reject hypothesis H_i having access only to the previous decisions. The original FDR online procedure guarantees control as soon as p-values are **independent**. However, our framework aims to set the confidence levels to the output levels α_j from the online FDR procedure which implies that **values used will not be independent** from previous hypothesis anymore. Thus, some modifications are necessary to embed Multi-Armed Bandit algorithms into an online FDR framework.

- Redefining null hypotheses : The first goal is to define a **null hypothesis** for each experiment. In fact, in our model we sample from multiple distributions adaptively, the null hypothesis could variate a lot between two experiments. Hence, given a distribution with default mean μ_0 and alternative distributions with means $\{\mu_i\}_{i=1}^{K}$, we can define the null hypothesis for the $j-th$ experiment as :

$$H_0^j : \mu_0 \geq \mu_i - \epsilon, \forall i = 1, ...K \tag{3}$$

Where μ_0 can be seen as $\mu_{control}$ which is the mean of the control arm.
As opposed to the null hypothesis H_0, we can define the false null hypothesis H_1 :

$$H_1^j : \mu_0 + \epsilon < \mu_i, \exists i \tag{4}$$

Thus, the null hypothesis states that the Modified Multi-Armed Bandit algorithm finds ϵ-better arm with confidence. Equivalently, there is no alternative arm that is ϵ-better than the control arm. This definition of the null hypothesis allows us to adapt to each experiment i. Intuitively, in order to confirm or infirm our null hypothesis, we need to define p-values for each experiment as well.

- Always-valid p-value : This kind of p-value already introduced in previous paper allows us to use a p-value at **arbitrary times** in the testing procedure and to **monitor the algorithm's progress in real time**. It can be defined as a stochastic process such that for all fixed and random stopping times T, under any distribution P_0 over the arm rewards such that the null hypothesis is true, we have :

$$P_0(p_t < \alpha) < \alpha \tag{5}$$

Hence, in simple words, whenever the "scientist" will stop the process, the null hypothesis will be guaranteed to be under the significance point α and thus will be significant. However, the initial definition of those always valid p-values only works for independent p-values. Furthermore, in our case, there is a dependence between p-values for each hypothesis j as the online procedure introduces a computed value of α_j directly influencing the next p-value at $j + 1$ and itself depending on rejection indicators of the null hypothesis, that are computed since the beginning of the process. Moreover, arms are pulled **adaptively**, which leads to the conclusion that the sample means are not **unbiased estimators** of the true means, since the number of times an arm was pulled now depends on the empirical means of all the arms. In order to deal with this issue, and be able to construct always valid p-values, we need to use the fact that p-values can be obtained by **inverting confidence intervals**.
Using the law of Iterated Logarithm : the below formula proves that any distribution P, with time t, alternative arm i with a mean μ_i, bounded by lower and upper confidence intervals, will be lower than a value γ, thus respecting the concept of **always validity**.

$$P(\exists t : \mu_i \in [LCB_i(t, \gamma), UCB_i(t, \gamma)]) \leq \gamma \tag{6}$$

With LCB and UCB being the lower and upper confidence bounds used in the best-arm algorithms. Those values can be computed as the average of rewards gained by an arm i up to time t ± a bouding function $\phi_n(\theta)$.
Knowing that, we can compute for each arm i at a time t for any fixed $\gamma \in (0, 1)$ the p-value as:

$$P_{i,t} = sup\{\gamma \in [0, 1] : LCB_i(t, \gamma) < UCB_0(t, \gamma)\} \tag{7}$$

The above equations shows that the p-value of a single arm can be computed by only by comparing the alternative arm's Lower Confidence bound with the control arm's upper confidence bound at instant t.
Then, we can compute the p-value at time t as the minimum of p-values over each of the K arms.

$$p_t = \underset{i=1,...K}{min} P_{i,t} \tag{8}$$

Finally, always valid p-value can be computed as the minimum p-value computed at time t and for each arm.

$$P_t = \underset{s\leq t}{min} p_t \tag{9}$$

We can intuitively understand those above formulas by defining a **sequential test T** which is a data-dependent rule for stopping the test and rejecting the null hypothesis:

- Stops the test later when α is lower
- Stops with probability $\leq \alpha$ when the null is true

Hence,

$$P_0(T_\alpha < \infty) \leq \alpha \tag{10}$$

Based on this definition, the p-value can be defined as the smallest α such that the α-level test would have stopped by observation n. The resulting value is **always valid**.
As p_n is decreasing, p_∞ exisit. Thus, for a fixed $\alpha \geq x$, the event $\{T_\alpha \leq \infty\}$ contains the event $\{p_\infty \leq x\}$. Thus, we can see that :

$$P_0(p_\infty < x) \leq P_0(T_\alpha < \infty) \leq \alpha \tag{11}$$

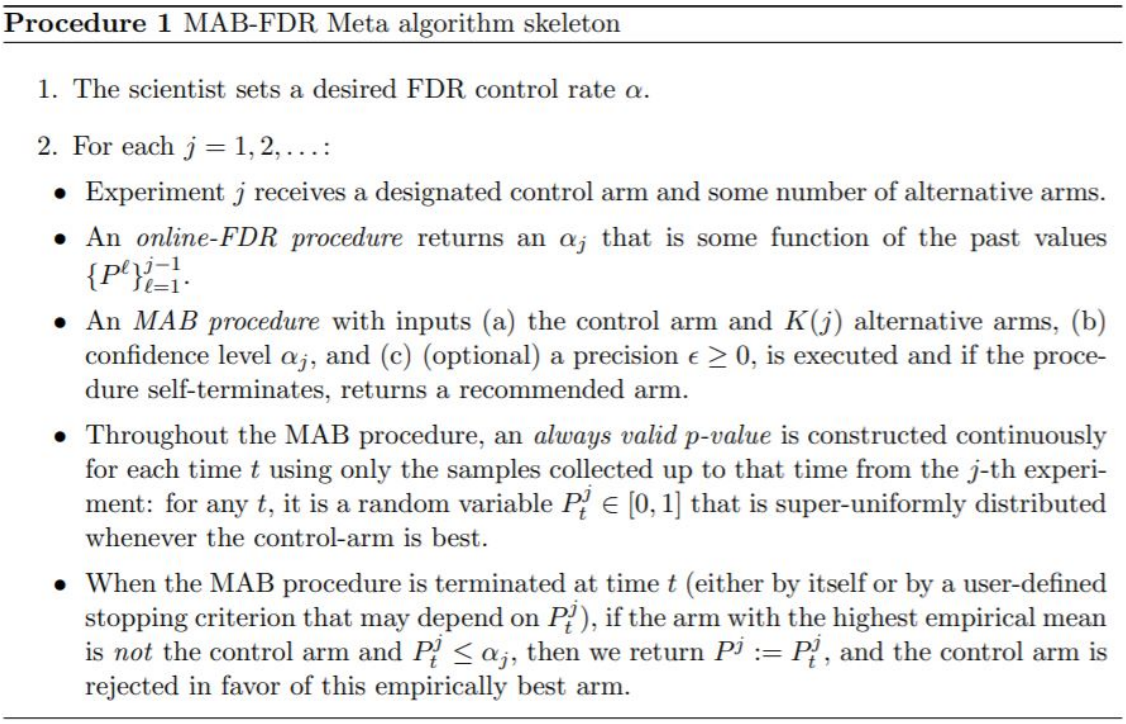
Thus, for any stopping time T :

$$P_0(p_T < x) \leq P_0(p_\infty < x) \leq x \tag{12}$$

As stated earlier, this construction of always valid p-values can be applied as our model is a doubly sequential procedure using Multi-Armed bandit algorithm. This allows us to apply continuous monitoring.
It is also important to notice that those above demonstrations are only valid for p-values which are conditionally **superuniform**. Which means that all of our p-values are stochastically dominated by a uniform random variable under the null.

2) Multi-Armed Bandit-online FDR

After having described key points that were needed to construct the framework, we can describe the overall procedure that allows the use of Multi-Armed Bandit within an online procedure.



Procedure 1 MAB-FDR Meta algorithm skeleton

1. The scientist sets a desired FDR control rate α.
2. For each $j = 1, 2, \dots$:
   - Experiment j receives a designated control arm and some number of alternative arms.
   - An *online-FDR procedure* returns an α_j that is some function of the past values $\{P^\ell\}_{\ell=1}^{j-1}$.
   - An *MAB procedure* with inputs (a) the control arm and $K(j)$ alternative arms, (b) confidence level α_j, and (c) (optional) a precision $\epsilon \geq 0$, is executed and if the procedure self-terminates, returns a recommended arm.
   - Throughout the MAB procedure, an *always valid p-value* is constructed continuously for each time t using only the samples collected up to that time from the j-th experiment: for any t, it is a random variable $P_t^j \in [0, 1]$ that is super-uniformly distributed whenever the control-arm is best.
   - When the MAB procedure is terminated at time t (either by itself or by a user-defined stopping criterion that may depend on P_t^j), if the arm with the highest empirical mean is *not* the control arm and $P_t^j \leq \alpha_j$, then we return $P^j := P_t^j$, and the control arm is rejected in favor of this empirically best arm.

In this procedure, the scientist has to introduce the first significant point α that will be the starting point of the computation the continuous computing of the next α for each hypothesis experiment j. Computing continuously the α at each experiment allows an overall adaptation of the procedure. This α_j depends itself on previous p-values computed by the Multi-Arms Bandit of the experiment. Each Multi-Arms Bandit j of the sequence will use this α_j in order to compute the always valid p-value and return a recommended arm. This clearly emphasis the dependency between experiments. Finally, in the end of the experiment, the last p-value computed will be compared to the significant level α and used to define whether or not to return the empirical best arm if not the control arm. According to this procedure, we can see that a Multi-Armed Bandit algorithm is used to achieve online FDR control along with **near-optimal sample complexity**.

Hence, intuitively comes a a way of taking advantage of the sample efficiency of best-arm bandit algorithms by setting the confidence levels close to what needed. Thus, our online FDR procedure is capable of carrying out significance levels α corresponding to the hypothesis for each experiment, given an "original" α initially set by the user. In this way, the FDR online procedure guarantees control **based on past decisions**.
To go further in the analysis, we can observe that our False Discovery Rate is at most α **at any time** of the experiment. Whereas if the algorithm is not terminated early, the power is at least $(1 - \alpha)$. Finally, we can observe that an appropriate bandit algorithm actually shapes the p-value distribution under the null in a good way that allows the control of the False Discovery Rate.

5 Best Arm Multi Armed Bandit

This paper introduces an alternative to the traditional best arm bandit algorithm which aims to deal with the introduction of asymmetry due to the adaptive null hypothesis seen earlier. Original best arm Multi Armed Bandit algorithm aims to identify the arm with the highest mean in a problem with high probability using a low number of samples. Thus, the supremum is taken over all set of means such that there exists an unique best arm.

Modified Best arm main particularities :

- checks if the current empirically best arm is within ϵ of the true highest mean, and if it is also at least ϵ greater than the true mean of the control arm (or is the control arm), terminates with this arm.
- ensures that the control arm is sufficiently sampled when $\epsilon \geq 0$
- pulls current empirically best arm and the most promising contender among the other arms, reducing the overall uncertainty in the difference between their two means

This algorithms shows typically the exploration/exploitation side of the Multi-Arm Bandit Algorithm. Hence, the goal is to find the empirical best arm while considering the fact that some arms have been more "pulled"/explored than other. Thus, the algorithm will pulls both the empirical best arm as well as the promising one to complete the probabilities needed to make the choice : reducing the uncertainty around the arms to be sure to make the right choice. In other words, this modified algorithm guarantees that when no alternative arm is ϵ-superior to the control arm, the algorithm stops and returns the control arm after a certain number of samples with probability at least $(1 - \gamma)$ (for any $\gamma \in (0, 1)$), where the sample complexity depends on ϵ-modified gaps between the means of the control arm and the alternative arm, taking into consideration the **asymmetry** between hypothesis. This also guarantees that if there is in fact at least one alternative that is ϵ-superior to the control arm, then the algorithm will find at least one of them that is at most ϵ-inferior to the best of all possible arms with the same sample complexity and probability.

6 Power Guarantee

The power of a binary hypothesis test is the probability that the test rejects the null hypothesis H_0 when a specific alternative hypothesis H_1 is true. In order to measure the power of the solution, the ϵ-best-arm discovery rate can be introduced.

$$\epsilon BDR(J) := \frac{\mathbb{E}\sum_{j\in H_1} R_j 1_{\mu_{ib}\geq\mu_{i*}-\epsilon} 1_{\mu_{ib}\geq\mu_0+\epsilon}}{|H_1(J)|} \tag{13}$$

where $H_1(J)$ represents the false null hypotheses up to experiment. $R_j = 1_{P_j \leq \alpha_j}$ indicates whether the null hypothesis of experiment j has been rejected. Hence,

$$R_j = \begin{cases} 1 & \text{if} P_j \leq \alpha_j, => \text{a claimed discovery that an alternative was better than the control} \\ 0 & \text{otherwise} \end{cases}$$

Moreover, this formula takes into consideration that the returned arm ib satisfies the bounds $\mu_{ib} \geq \mu_i - \epsilon$ as well as $\mu_{ib} \geq \mu_{i*} - \epsilon$ with the corresponding indicator functions added to the numerator. This is intuitive as we want the mean of our new arm to be within ϵ of the mean of the best arm μ_* as well as at least ϵ-better than the control arm with mean μ_0.
Using this metric, the proof of the power using a procedure of best arm multi-armed Bandit algorithm with online FDR procedure named **LORD** shows that the ϵ-BDR is lower bounded when the maximal number of samples the scientist wants to pull is infinite.

$$\epsilon BDR(J) \geq \frac{\sum_{j=1}^{J} 1_{j \in H_1}(1-\alpha_j)}{|H_1(J)|} \tag{14}$$

Hence, the power of the solution is guaranteed as the rate of ϵ-best arm discovery is guaranteed to be higher than a value directly depending on our specific alternative hypothesis at experiment j, $H_1(J)$ and the significance level of the same experiment α_j.

7 Analysis

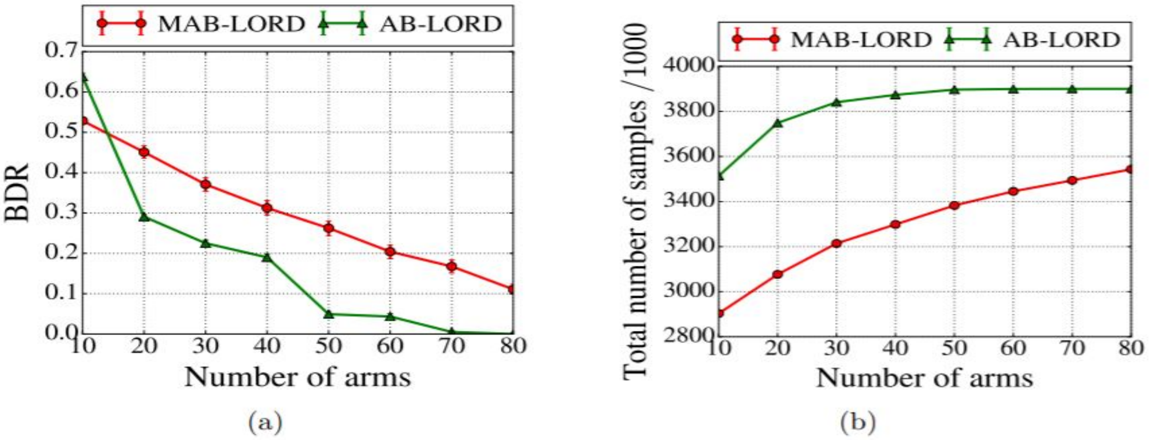
We have seen that this frameworks combines already known procedures in order to reach its objectives which are :

- Adaptive Sampling using Best-arm multi-armed Bandit that allows many variants in the test (represented by the arms)
- Construct always valid p-values that allows p-value control.
- Online FDR control procedures that allows multiple testing over time.

The introduced procedure is defined as sequential, and simultaneously under certain conditions can:

- Guarantee a near optimal sample complexity : the total number of necessary 'pulls' to determine significance is optimal.
- Yield Near-optimal best-alternative discovery rate (high power)
- Controls FDR in an online fashion (low FDR at any point in time)

A condition for this procedure to be respected is the superuniformity of p-values.



The truncation time is random stopping time for the experiment.

The above experimental results show how the introduced procedure yields better ϵ best arm discovery rate than the A/B test using the same online procedure (LORD) for both of them. In fact, as expected, the power of the solutions is shown to be be better for MAB-LORD than AB-LORD. However, the gap between the two results' BDR is quite low (0.1) which could question the choice of the analysed data which could yield more significant results. In another hand, the number of arm pulled by the MAB-LORD procedure is always optimal compared to AB-LORD.
Hence, those results are really interesting for a real life scenario, for instance, when patients are assigned a treatment. The fact that less arms need to be 'pulled' while increasing the number of arms/options; or the need for less exploration with the proposed framework than with the A/B testing shows that less patients would suffer from trying drugs that were not the best options. Evenly in a marketing context, the gain

would be immediate compared to A/B testing as the exploration would be much less systematic whereas the exploitation, which ensure gain, would be faster.
Hence, this framework can be used to address different issues with good guarantees of efficiency and quality.
As stated in the paper, the results can be extrapolate to other online FDR procedures as well as other best-arm multi-armed Bandit algorithm, the efficiency of the framework will mainly depend on the efficiency of either of those two. Thus, it would be interesting to try using several multi-armed bandit algorithm than LUCB.

8 Questioning and limitations

Extrapolation of results :

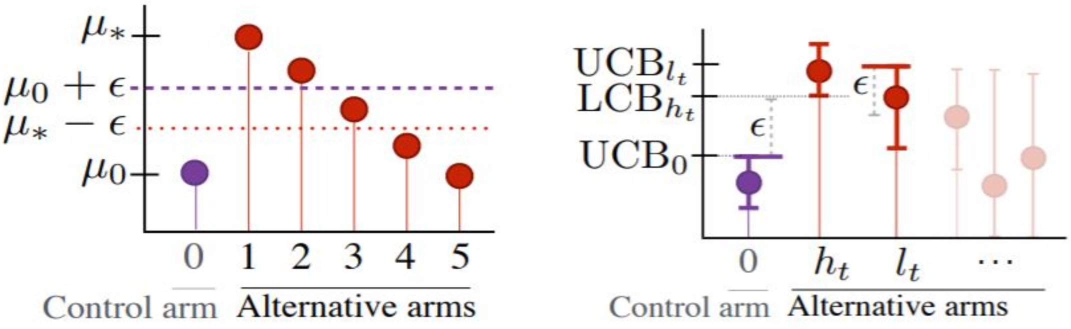
This paper showed guarantees of :

- Yielding good sample complexity
- Allowing continuous monitoring
- Controling FDR onine

However, the guarantee of FDR and mFDR control are only proven with the use of the **MAB-LORD** or a **modified MAB-LORD** algorithm. As stated in the paper, this framework is meant to be used with a set of online False Discovery Rate procedures as well as best-arm multi-armed Bandit algorithms that are defined by the scientist using it. Hence, there is no explicit nor proven guarantee that this framework works concretely with other procedures. Moreover, the experimental results are only showing that the MAB-LORD procedure, is better than the AB-LORD procedure, but what about other online False Discovery Rate procedure that could be necessary to other concrete applications ? Furthermore, how much change needs to be done in existing online procedure as well as new FDR procedure in order to get those results ? A more general approach during the experiment (using for example several kind of online FDR procedures) could have been useful to underline/prove the flexibility of this framework.

Use of the slack variable ϵ :

While introducing the algorithms and the different metrics such as the ϵBDR we take into account the slack variable ϵ. However, this one is never used in the different experimental test : set to 0. This choice from the author is seen as a matter of simplicity to focus on more important results. However, this variable is used all over the paper to introduce bounding and control in the procedures.
For example, the multi-armeds Bandit algorithm uses this variable which affects the **stopping criterion** as well as the confidence bounds as shown in the following figures.



(a) (b)

Hence, figure above shows how ϵ affects the choose of the best arm. In fact, according to the plot (a), the mean of arms 1, 2 and 3 are within ϵ of the best arm but only arms 1 and 2 are at least ϵ better than the control arm with mean μ_0. Returning 3,4 or 5 would have been a false discovery for $\epsilon 0$ Hence, setting ϵ to 0 would change the results as only the arm 1 would have been taken into account within the best-arm whereas arms 1,2,3,4 are at least 0 better than the control arm. The same analysis can be made with the plot b where the ϵ criterion directly affects the stopping condition where $LCB_{empiricalbestarm} \geq UCB_{promisingalternativearm} + \epsilon$ Thus, we believe that setting it to 0 facilitates the experiment and doesn't show the true potential of the framework. The reader could believe that setting this ϵ to a non-null value would affect the performances of the experiments and could yield more ”close to reality” results.

Trade-off guarantee :

As stated in this paper, the small value of α guarantees smaller False Discovery Rate error as well as a higher best arm discovery rate, guaranteeing power and accuracy of the model. However, we have also seen that α at each experiment j (α_j) are directly dependent on the initial α set by the user. Hence, a smaller initial α set by the scientist will imply a smaller α_j at each experiment. Thus, the multi-armed Bandit algorithm used at experiment j, which uses α_j in order to construct $p - values$ will need to employ a larger number of ”pulls” in each experiment. This will directly affect the sample-optimality guaranteed in the paper. Thus, trade-off must be taken into account by the scientist and a good-understanding of the importance of power and accuracy against sample optimality is necessary to make a good use of the framework.

Limitation of online False Discovery Rate :

This paper uses LORD procedure as online False Discovery Rate algorithm. LORD procedure is also described as a kind of ”generalized alpha investing rule”.
However, in their paper ”Online Rules for Control of False Discovery Rate and False Discovery Exceedance” [1] in which they introduce this procedure, Javanmard and Montanari, warn about the use of the LORD procedure. According to them, extra caution should be taken when the false discovery proportion can deviate significantly from its expectation, represented in our case as the False Discovery Rate. This can occur whenever:

- The number of hypothesis is not very large.
- There is a significant correlation between them.

However, guarantees and procedures shown in the paper using multi-armed Bandit together with LORD procedure while experimenting doesn't take into account those warnings which could have a harmful impact in a real case situation.
A first thought suggestion could be to use the false discovery exceedance instead of the False Discovery Rate, and imposing additional constraints on generalized alpha investing rules such as LORD. The **False discovery exceedance** FDR can be viewed as the expectation of false discovery proportion. In fact, in some cases, False discovery proportion may not be well represented by its expectation, for example, when the number of discoveries is small. In these cases, the false discovery proportion might be larger than its expectation with significant probability. In order to provide tighter control, we develop bounds on the false discovery exceedance. The author should thus be aware of the limitation of online False Discovery Rate procedures that could affect its results under some circumstances.

Bias in publications :

One limitation that is really common into research paper is the introduction of bias. In fact, the effect of bias can directly increase the False Discovery Rate. There are many sources of bias, but in our paper only few of them can be targeted as potentially affecting the model.

- **Flexibility in the design :** As stated before, the experiment of multi-armed bandit algorithm offers a lot of flexibility in terms of parameters, design of the framework, outcomes, and possible analysis. Thus, there is a huge potential of inducing negatives results into positive ones. Knowing that, a suggestion would be to ”standardize” the methodology, the way of approaching the given problem which could be likely to reduce the risk of false positives. In our case, proposing clear methods to adopt with the framework as well as suggestions or limits for the tuning parameters used in the model would be an option.
- **Pre-selection** : The more hypothesis are tested, the more likely it is to find false positives. This can apply to our model which uses sequences of hypothesis that can be infinite.

Thus, if introduced, this bias could be harmful to the experimental results showed in this paper. It is important to prevent it as much as possible.

9 To go further

Some other approach could be relevant to this framework. First of all, the introduction of **Quality Preserving Database** in the paper ”Novel Statistical Tools for Management of Public Databases Facilitate Community-Wide Replicability and Control of False Discovery”[6] from Saharon Rosset, and Ehud Aharoni could be interesting to take into account as future evolution in the framework. The Quality Preserving Database is presented an approach to public database management that enables perpetual use of the database for testing statistical hypotheses while controlling false discovery and avoiding publication bias on the one hand, and maintaining testing power on the other hand. More precisely, it is a database with a management layer that assigns costs in the form of additional data samples for each test executed. This layer fulfills properties:

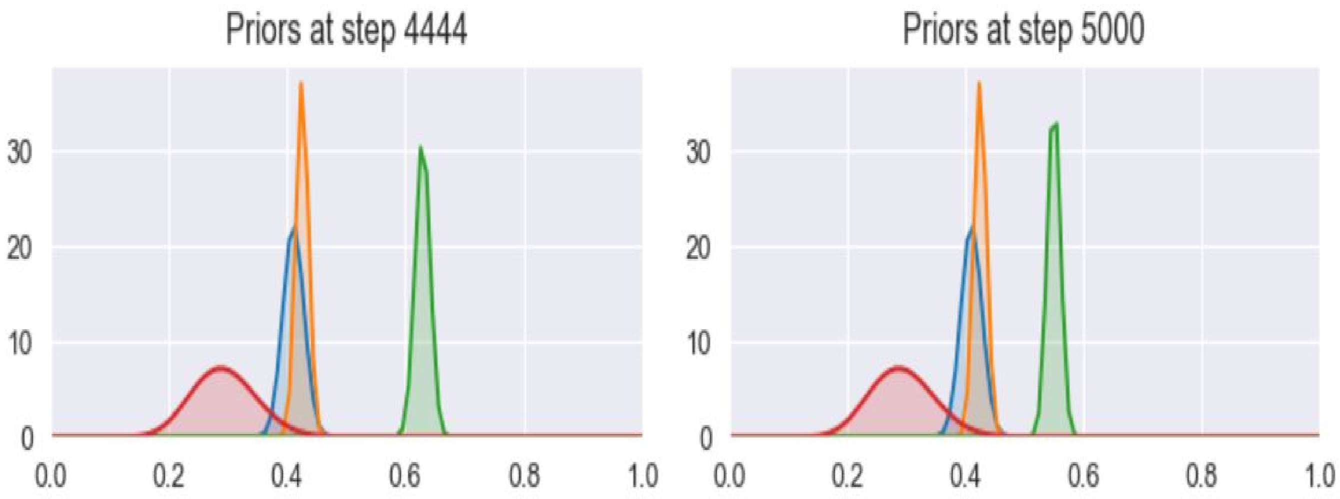
- It can serve an infinite series of requests
- It satisfies the fairness and stability requirements,
- It controls some measure of the overall type-I errors (false positive) at some pre-configured level α.

What if we could use the multi-armed bandit algorithm in a online fashion with the Quality Preserving Database. Will the drawbacks of FDR and bias be corrected thanks to this hybrid method ? Will it introduce new issues that we didn't foresee ?

Knowing that each arms of the multi-armed Bandit algorithm is univariate, extending the framework to less specific settings would be an other opening.
Thus, using results of the paper ”Sequential nonparametric testing with the law of the iterated logarithm”[2] from A. Balsubramani and A. Ramdas, which shows how to construct sequential tests for many multivariate non parametric testing problems, using Law of iterated logarithms confidence intervals, which can again be inverted to provide always valid p-values, introduced earlier. A new, more complete hybrid framework could arise from this work. Using this idea, the main idea is to set the null hypothesis that the control arm has the same mean as other arms. Hence, picking the arm from which the mean is furthest away from the control could be the alternative hypothesis.
Furthermore, using pairs of arms which would induce dependencies and setting the null hypothesis as the ”rewards in the control arm are independent of the alternative”, we could pick the most correlated arm if the null hypothesis was rejected. Finally, the framework could be modified to fit the the contextual bandit setting in which the samples are associated with features to aid exploration instead of treating samples as identical from a statistical perspective.

10 Numerical exploration

After having seen the introduction of a framework smartly combining multi-armed Bandit algorithm together with online False Discovery Rate control, we decided to investigate further into the analysis of the best arm - Multi arm bandit algorithm. We used Thompson sampling which is known for yielding good results, especially in marginal cases. Hence, we could imagine combining the Thomson sampling together with the online False Discovery Rate control into the framework presented previously. It would be interesting to compare the results between the Thomson Sampling and the LUCB best arm algorithm in order to analyse differences and tackle out some uncertainty highlighted as limitations. The analysis of the MAB-Thompson and its explanations are provided in the appendix: ”theoritical guidelines project mallet gajendran”.



Taken from the numerical application, the figure above shows the distribution of each arm being the best alternative. The green arm has grown while the others remained stable

References

[1] Andrea Montanari Adel Javanmard. ”online rules for control of false discovery rate and false discovery exceedance”.

[2] A. Balsubramani and A. Ramdas. ”sequential nonparametric testing with the law of the iterated logarithm”. In *”Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence”*, page . 42–51, 2016.

[3] Dallas Card and Shashank Srivastava. ”summary and discussion of: ”why most published research findings are false”. In *Statistics Journal Club,*, pages 36–825, 2014.

[4] Kevin Jamieson Martin J. Wainwright Fanny Yang, Aaditya Ramdas. ”a framework for multi-a(rmed)/b(andit) testing with online fdr control”.

[5] Sivan Sabato. ”ϵ-best-arm identification in pay-per-reward multi-armed bandits”.

[6] Ehud Aharoni Saharon Rosset. ”novel statistical tools for management of public databases facilitate community-wide replicability and control of false discovery”.

[5] [3]

https://towardsdatascience.com/comparing-multi-armed-bandit-algorithms-on-marketing-use-cases-8de62a851831