On the Advances and Challenges of Adaptive Online Testing

Bo Yang*
LinkedIn Corporation
Sunnyvale, California, USA
by9ru@virginia.edu

Da Xu*
Walmart Labs
Sunnyvale, California, USA
daxu5180@gmail.com

ABSTRACT

In recent years, the interest in developing adaptive solutions for online testing has grown significantly in the industry. While the advances related to this relative new technology have been developed in multiple domains, it lacks in the literature a systematic and complete treatment of the procedure that involves exploration, inference, and analysis. This short paper aims to develop a comprehensive understanding of adaptive online testing, including various building blocks and analytical results. We also address the latest developments, research directions, and challenges that have been less mentioned in the literature.

KEYWORDS

Adaptive decision making, Online exploration, Confidence bound method, Hypothesis testing, Information theory

1 INTRODUCTION

Decision making in modern industry is often driven by randomized experiments behind the scenes. In particular for IT companies, the production scenario ranges from deciding the best-performing recommendation system to determining the size of an ad banner. The traditional process of simultaneously launching alternative services to randomly assigned users is also known as *online A/B/n testing* [12, 25]. While A/B/n testing and its variants enjoy a high degree of simplicity in planning and making inference, they still lack adaptability when certain alternatives become clearly inferior. Perhaps the most critical consequence would lost revenue and user satisfaction, so optimizing online testing strategy has become imperative for many, if not all.

We start our discussion by first characterizing an online experimentation. By convention, we use the notion of *arm* to indicate such as a candidate webpage service, denote by: $a \in \mathcal{A}$ with $|\mathcal{A}| = k + 1$. Sampling (obtaining) an observation from arm a is represented by: $X_{a,i} \sim P_a, i \in \{1, \ldots, n_a\}$, where P_a is the reward (e.g. impression rate, revenue, return of investment) distribution associated with arm a. We use a_0 denote the *control arm* which often refers to the currently deployed service. We will follow this convention unless specified. While $X_{a,i}$ are often assumed to be identically distributed under each arm a, they are not necessarily independent. Handling dependency will be an important topic in our following discussion, since it cannot be avoided by adaptive exploration, but would violate traditional statistics inference assumptions [8].

For the most part, the inference goal of an online experiment is to test a statistical hypothesis often related to the data distribution, e.g. whether or not $\mu_1 > \mu_2$ where $\mu_q := \mathbb{E}[X_{a_q,i}]$ is the expected reward for arm q. The outcome of a hypothesis testing is usually expressed via the null hypothesis H_0, alternative hypothesis H_1, a significance level $\alpha \in (0, 1)$, and the observed samples. For example, when H_0 is rejected, it holds:

$$Pr\big(\text{rejecting } H_0 \text{ for } H_1 \mid H_0 \text{ is true}\big) \leq \alpha, \tag{1}$$

In other words, it is unlikely that H_1 better characterizes the observed data distribution just by chance. Also, the control arm a_0 usually plays a pivotal role in the hypothesis's design: a_0 is often assumed to be superior unless overturned by enough evidence.

Practically for IT companies, adaptive online testing means the proportion of traffic guided to each arm during the random assignment can be adjusted based on the performance to date. Since the purpose of online testing is purely *exploratory*, i.e. to explore as much as possible without worrying about the cost, the problem setting differs naturally from *bandit* or *reinforcement learning* where exploitation is mandatory [5, 31]. Nevertheless, some exploration strategies from the bandit literature also suit the pure exploration task [26], and they are often studied under the *best-arm identification* problem [4, 27].

In particular, suppose $a^{[i]}$ is the arm with the ith highest reward, and denote by: $\Delta_j = \mu_{a^{[1]}} - \mu_{a_j}$ which is the *suboptimality gap* of arm a_j. We use $\mathbf{h}_T$ to represent the first T samples where $\mathbf{h}_T \in \mathcal{H}_T$. The exploratory algorithms from best-arm identification can be described by a stopping time[1] T and a data-adaptive decision rule $\mathbb{D} : \mathcal{H}_T \to [k]$ such that when the algorithm halts, the probability that $a^{[1]}$ is not identified (which equivalents to the event of $\Delta_{\mathbb{D}(\mathbf{h}_T)} > 0$) becomes small:

$$Pr\big(T < \infty \text{ and } \Delta_{\mathbb{D}(\mathbf{h}_T)} > 0\big) \leq \delta, \tag{2}$$

where $\delta \in (0, 1)$ is also some pre-defined confidence level.

While best-arm identification and hypothesis testing are similar in finding the best candidate (hypothesis or arm) with some tolerance of making mistake due to randomness, their meanings are ultimately different. One apparent reason is that the probability in (2) is taken w.r.t. the sampling and reward distributions, while the conditional probability in (1) is taken w.r.t. the distribution under H_0. The conclusion of one problem does not directly imply another. Meanwhile, notice that as long as the correct statistical procedure is employed, the hypothesis testing does not necessarily depend on the sampling strategy. There is possibility to employ the sampling strategy from best-arm identification and use hypothesis testing as part of the stopping rule. This integration may bring adaptability to online testing while not costing its statistical rigorousness. Now both problems will be subject to new changes:

---

[1]We assume the measurability of T is granted throughout our discussion.

- the current sample may depend on the past ones due to the adaptive sampling strategy;
- there will be continuous interactions between the stopping criteria (and stopping time) of the best-arm identification, the p-value, and the significance level of hypothesis testing.

The first change rules out traditional hypothesis testing procedures (such as the *Student's t-test*) since the observations are no longer i.i.d. The impact of the second change is more involved since it concerns the working mechanisms of both best-arm identification and hypothesis testing. The good news is we can leverage *confidence bounds* to connect the decision-makings of best-arm identification with hypothesis testing. Nevertheless, the nature of p-value and significance level will be different in the sequential setting, and they need to be adjusted along the path of adaptive exploration. Finally, even the best strategy may struggle to find the best-performing arm if the structure of a particular problem instance is unfriendly. Studying this *instance-wise* difficulty would be valuable for understanding the inherent challenges of adaptive online testing.

Section 2 covers the basics of pure exploration and hypothesis testing with a highlight on confidence interval. We study in Section 3 the more advanced tools and solutions for integrating interact hypothesis testing with best-arm identification. In Section 4, we discuss the difficulty of best-arm identification and the challenges for adaptive online testing. We may omit technical details for the sake of space, and readers may refer to the references provided.

Contribution: since our paper is the first in IR literature that systematically discuss exploration, inference, and analysis of adaptive online testing, it aims to serve both as introductory material for unfamiliar readers and a technical summary for domain experts.

2 PURE EXPLORATION, CONFIDENCE INTERVAL, AND P-VALUE

The best-arm identification in (2), which we described as a pure exploration problem, takes the *fixed confidence* setup[2] [15]. Using exploration strategy to optimize online problems means finding the optimal arm with as few samples as possible. In particular, the strategy must be carefully designed so each arm is explored just enough to be separated from the others with high confidence. The traditional A/B/n testing acts like a *uniform exploration* policy where the stopping rule is to reach the maximum round of experiments or reject the null hypothesis. Probably the most important idea behind many pure exploration algorithms is the *confidence bound* (CB) method [5], where empirical uncertainty is used to construct optimistic/pessimistic overestimations of the mean reward.

Note that the randomness in finite samples can cause any estimation of $\{\mu_i\}_{i=1}^k$ subject to uncertainty. Under mild distribution assumptions such as P_a is *Gaussian* or *sub-Gaussian*, there exists a (provably tight) region that contains μ_i with arbitrarily high probability. The region is often constructed using finite-sample concentration results such as the *Hoeffding* or *Chernoff* inequalities [7], which we exemplify as below.

LEMMA 1 (CONCENTRATION OF I.I.D. MEAN). *Assume* $X_{a,i}, a = 0, \ldots, k$ *are independent,* σ-*sub-Gaussian random variables. Then for*

any $\epsilon > 0$:

$$Pr\left(\frac{1}{n_a}\sum_{i=1}^{n_a} X_{a,i} - \mu_a \geq \epsilon\right) \leq \exp\left(\frac{-n_a\epsilon^2}{2\sigma^2}\right), \quad (3)$$

and the same bound holds for $Pr(\frac{1}{n_a}\sum_{i=1}^{n_a} X_{a,i} - \mu_a \leq -\epsilon)$.

We immediately observe that by solving for $\exp\left(\frac{-n_a\epsilon^2}{2\sigma^2}\right) = \frac{\delta}{2}$, it holds with probability at least $1 - \delta$:

$$\mu_a \in \underbrace{\left[\hat{\mu}_a - \sqrt{2\sigma^2 \log(1/\delta)/n_a}\right.}_{\text{LCB}^{(a)}(n_a,\delta)}, \underbrace{\left.\hat{\mu}_a + \sqrt{2\sigma^2 \log(1/\delta)/n_a}\right]}_{\text{UCB}^{(a)}(n_a,\delta)}, \quad (4)$$

where $\hat{\mu}_a = 1/n_a \cdot \sum_{i=1}^{n_a} x_{a,i}$, and *LCB* and *UBC* are shorthands for the lower/upper confidence bound. Both *UCB* and *LCB* express as: $\hat{\mu} \pm \text{CI}(n_a, \delta)$, where *CI* is the confidence interval that depends on the sample size and confidence level δ.

For any two arms a_1 and a_2, $\text{LCB}^{(a_1)}$ and $\text{UCB}^{(a_2)}$ provide the pessimistic and optimistic estimation of μ_1 and μ_2, respectively. If $\text{LCB}^{(a_1)} > \text{UCB}^{(a_2)}$, it is easy to show that $\mu_1 > \mu_2$ with high probability, again by using concentration inequalities. Therefore, the CB method provides an elegant but powerful approach to tailor decision making based on the uncertainty of an estimation. The two major algorithms for best-arm identification, namely *Action Elimination* and *Upper Confidence Bound*, both rely on CB:

- **Action Elimination** [11]: sample from each arm a pre-determined n_{init} times, and eliminate according to: $\text{UCB}^{(a_i)} < \text{LCB}^{(a_0)}$ where a_0 is a reference arm. Continue until only one arm is left.
- **Upper Confidence Bound** [4, 17]: keep sampling from the arm indexed by $\arg\max_a \text{UCB}^{(a)}$, and stop when there exists an a^* such that: $\text{LCB}^{(a^*)} \geq \text{UCB}^{(a_i)}$ for all $i = 1, \ldots, k$.

We point out that both approaches have their limitations when used as an exploration strategy for online testing. The merit of Action Elimination is to throw away arms only after we accumulate enough confidence on their suboptimality. Although it improves upon the uniform exploration, the suboptimal arms is still likely to be explored more than necessary. The Upper Confidence Bound is on the opposite side: the leading arm will be overexplored, and even the second-best arm may be underexplored. Luckily, both of them can be improved by using an variant of the Upper Confidence Bound method, which is known as *LUCB*. LUCB aims to find a better balance between exploring the leading arm and others:

- **LUCB** [20, 23]: keep sampling from both the current *best* and *second-best* arm (denote by a^* and a^{**}) in terms of $\{UCB^{(i)}\}_{i=1}^k$, and stop when: $\text{LCB}^{(a^*)} > \text{UCB}^{(a^{**})}$.

For all the above algorithms, confidence intervals not only guides how to explore, but also decides when to stop. As we show next, this will be an advantage for bringing in hypothesis testing, because confidence intervals can also be related to p-value.

Although criticized for biasing scientists and researchers toward a particular perspective of experimentation [3], *p-value* remains an important metric to understand the significance of a discovery. Generally speaking, p-value is the probability of obtaining test results at least as extreme as what is observed, if the null hypothesis H_0 is correct. It usually involves a *test statistic Z* whose distribution is known under H_0. Suppose the observed test statistic is z, then

---

[2]There are two related but different settings in best-arm identification: fixed confidence and fixed budget. The solutions can often be switched between the two settings [14], so we stick with the fixed confidence problem which has been studied in more depth.

the probability of having more extreme value on both sides of the distribution is given by: $Pr(|Z > z| \,|\, H_0)$.

When the quantity of interest is mean reward, one straightforward choice is letting: $Z_k := \frac{1}{n_0} \sum_{i=1}^{n_0} X_{a_0,i} - \frac{1}{n_k} \sum_{i=1}^{n_k} X_{a_k,i}$. Under the null hypothesis $H_0 : \mu_{a_0} = \mu_{a_k}$, for independent (sub)-Gaussian X_i, the test statistic Z_k follows a *zero-mean* (sub)-Gaussian distribution. If the distribution of Z_k is symmetric, the p-value is:

$$Pr(|Z_k > z_k| \,|\, H_0) = Pr\left(\left|\frac{1}{n_0} \sum_{i=1}^{n_0} X_{a_0,i} - \frac{1}{n_k} \sum_{i=1}^{n_k} X_{a_k,i}\right| > |z_k|\right).$$

As an immediate result, p-values can also be obtained from concentration inequalities similar to the ones in Lemma 1, since they both measure the deviations under some fixed distribution. Following the same logic as CB, it is easy to check that $Pr(Z_k > z_k \,|\, H_0) \leq \alpha$ (which α is the pre-defined significance level) also leads to a confidence-interval-based decision making, e.g.:

$$\text{reject } H_0 \text{ if } z_k \notin \left[-\text{CB}(n_k, n_0, \alpha), \text{CB}(n_k, n_0, \alpha)\right],$$

where CB is the associated confidence interval whose expression we omit here for brevity. In fact, interested readers may check that the single-arm p-values for comparing a_0 and a_k under $H_0 : \mu_{a_0} \geq \mu_{a_k}$ can be obtained by inverting *LCB* and *UCB* via [24]:

$$P_{n_a,n_k} = \sup\left\{\gamma \in [0,1] \,\big|\, \text{LCB}^{(a_k)}(n_k, \gamma) \geq \text{UCB}^{(a_0)}(n_0, \gamma)\right\}. \quad (5)$$

The above link between p-value and confidence interval demonstrate the feasibility of combining best-arm identification with hypothesis testing. In particular, there could exist a unified decision-making procedure using only confidence intervals. However, there are several obstacles that require further investigation, notably:

(1) how to obtain valid confidence intervals when the observations are no longer independent;
(2) how to make sure the p-value is always valid and the false discovery rate is controlled, since the testing can be queried by the adaptive exploration algorithm at any time.

In the next section, we discuss some of the more advanced tools and methodologies that may address the above challenges.

3 ADVANCES IN BEST-ARM IDENTIFICATION AND HYPOTHESIS TESTING

When X_i are not independent, or when the sample size $n(T)$ itself is a random variable that depends on some stopping time, $\sum_{i=1}^{n} X_{a,i}$ becomes a *random walk* whose concentration behavior differs from the Hoeffding-type inequalities we introduced earlier. Historically, the *law of iterated logarithm* (LIL) [16] was developed to describe the asymptotic behavior of a zero-mean unit-variance random walk:

$$\limsup_{n \to \infty} \frac{\sum_{i=1}^{n} X_i}{\sqrt{2n \log \log n}} = 1, \quad \text{almost surely.} \quad (6)$$

Unfortunately, LIL only characterizes the limiting behavior, while real-world online testings often falls into the finite-sample regime. The recent work by Zhao et al. [33] propose a novel finite-sample bound for such cases, from which we can construct confidence intervals for both best-arm identification and hypothesis testing.

LEMMA 2 (CONCENTRATION OF RANDOM WALK). *Let T be any random variable taking value in $\mathbb{N}$. Let $S_T := \sum_{i=1}^{T} X_i$. There exists*

$a > 0$ *such that:*

$$Pr\left(S_T \geq \sqrt{0.6n \log(\log_{1.1} n + 1) + an}\right) \leq 12 \exp(-1.8a). \quad (7)$$

Comparing the convergence rates implied by Lemma 1 and 2, we spot a $\sqrt{\log \log n}$ difference caused by the extra stochasticity from dependent observations. This also holds asymptotically if we compare LIL with the *law of large numbers*. As a result, the confidence intervals for adaptive online testing may just be wider than its non-adaptive counterpart. It means that under the same context, more observations may be required to trigger a decision.

Now that we know how to compute confidence bounds correctly, the next challenge is to make p-values *valid at all times*. This step is critical if the stopping rule of online exploration can consult the p-value at any time. Just like having a monkey randomly tap a typewriter indefinitely, it will almost surely enter any given text. Ideally, an any-time p-value should allow guilt-free queries from the continuing adaptive exploration strategy, and the previous study by Johari et al. [19] suggests that a always valid p-value is a stochastic process $\{P^{(t)}\}_{t=1}^{\infty}$ such that for any (random) stopping time T, under any distribution $\mathbb{P}_{\mathcal{H}_0}$ that agrees with the null hypothesis, we have:

$$\mathbb{P}_{\mathcal{H}_0}(P^{(T)} \leq \alpha) \leq \alpha. \quad (8)$$

Note that this definition extends a classical property of p-value that it should follows *uniform distribution* under $\mathbb{P}_{\mathcal{H}_0}$.

Recently, Yang et al. [32] proposes taking the *minimum* of all single-arm p-value, from all the previous steps, to obtain an *any-time* p-value such that it satisfies (8). For instance, at step T, let $n_0(T)$ and $n_k(T)$ be the #samples from arm a_0 and a_k collected until the Tth step, then the any-time p-value for the experiment is:

$$P_{a_0,a_k}^{(T)} := \min_{s \leq T} \min_{i=1,\ldots,k} P_{n_0(T),n_i(T)}.$$

The idea being taking all the minimums is to make sure that: $\mathbb{P}\left(\bigcup_{j=1}^{k} \bigcup_{t=1}^{\infty} \left\{P_{n_0(t),n_j(t)} \leq \alpha\right\}\right) \leq \alpha$, so $P^{(T)}$ is always valid.

It is also known that controlling the false alarm error of each test is not sufficient to achieve a small *false discovery rate* (FDR) when multiple testings occur. Let $\alpha^{(t)}$ be the significance level at step t. When running an adaptive online testing, $\mathbb{1}\left[P^{(t)} < \alpha^{(t)}\right]$ will be queried multiple times. Therefore, to control FDR in an online fashion (we use $R^{(t)} = 1$ to denote claiming a discovery at step t), it is intuitive that $\alpha^{(t)}$ should be aware of the history $\{R^{(j)}\}_{j=1}^{(t-1)}$:

$$\alpha^{(t)} := \alpha^{(t)}\left(\alpha; R^{(1)}, \ldots, R^{(t-1)}\right),$$

by using a proper decision rule $\alpha^{(t)}$. Formally, the FDR at step t is defined as $\mathbb{E}\left[\frac{\text{#True discoveries}}{\text{#Discoveries made before } t}\right]$. In the non-adaptive setting, the *Benjamini-Hochberg* procedure is a standard way to control FDR [6]. For adaptive experimentation, we find the idea of "α-investing" powerful for controlling online FDR [13]. The general idea is to let the decision rule $\alpha^{(t)}$ "invest" a small amount of confidence each time when a testing is called, and $\alpha^{(t)}$ will gain some reward if a discovery is claimed. By choosing the proper investment and reward function, this procedure has been shown to achieve FDR control. The implementation is quite straightforward and readers may refer to the references herein [2, 18, 28].

We now put together the pieces and present an exemplary adaptive online testing procedure. We first sample from each arm n_{init} times for warm-start. After that, at each step t:

(1) Obtain the adjusted significance level $\alpha^{(t)}$ using the history p-values $\{P^{(j)}\}_{j=1}^{t-1}$ and rejections $\{R^{(j)}\}_{j=1}^{t-1}$, with α-investing.

(2) First follow LCUB: compute $\{\text{UCB}^{a_i}(n_i(t), \alpha^{(t)})\}_{i=0}^{k}$, find a^* and a^{**}. If not LCB$^{a^*}$ > UCB$^{a^{**}}$ and LCB$^{a^*}$ > UCB^{a_0}, sample from a^* and a^{**}; otherwise return them. Then compute and track the any-time p-value $P^{(t)}$.

(3) Record $R^{(t)} = \mathbb{1}[P^{(t)} > \alpha^{(t)}]$, and decide whether to stop based on $R^{(t)}$ or other stopping criteria.

4 CHALLENGES OF ADAPTIVE ONLINE TESTING

Although the recent advancements can refine the procedure of adaptive online testing, they do not change the nature of the problem, especially its inherent difficulty. Here, the difficulty refers to the minimum number of samples required by whatever smart algorithm to achieve high-confidence decisions. For simplicity, assuming that each arm has unit variance. For uniform exploration, it is obvious that the difficulty lies exactly in how separable the set of arms are (in terms of μ_a). Adaptive explorations may help us eliminate inferior arms faster, but the optimality of the eventual performance will depend more heavily on the gap between $a^{[1]}$ and $a^{[2]}$. In general, the distribution patterns of the suboptimality gaps can induce different levels of difficulty for exploration, and the recently Chen et al. [9] reveals a (tight) lower bound to justify this phenomenon.

In the theoretical literature, information-theoretical lower bound if often used to describe the amount of information needed to specify the answer related a particular problem instance. In the case of best-arm identification, the problem instance can be effectively described by the set: $\mathcal{I} = \{\Delta_{[j]}\}_{j=2}^{k}$, where $\Delta_{[j]} := \mu_{a^{[1]}} - \mu_{a^{[j]}}$.

LEMMA 3 (DIFFICULTY OF BEST-ARM IDENTIFICATION). *Let* $T_{\mathbb{D}}(\mathcal{I}, \delta)$ *be the stopping time for applying* $\mathbb{D}$ *on problem* $\mathcal{I}$ *such that it achieves the δ-correctness defined in (2). It holds that:*

$$\inf_{\mathbb{D}} T_{\mathbb{D}}(\mathcal{I}, \delta) = O\Big(\sum_{j=2}^{k} \frac{1}{\Delta_{[j]}^2} \Big(ln\frac{1}{\delta} + H(\mathcal{I}) \Big) + \frac{ln\,ln\,\Delta_{[2]}^{-1}}{\Delta_{[2]}^2} polylog(k, \delta^{-1}) \Big),$$

where $H(\mathcal{I})$ is an entropy term of the set $\mathcal{I}$.

Before we discuss this result, recall that adaptive exploration does not account for the asymmetry of the null and alternative hypothesis. Yang et al. [32] recently propose to give the control arm a little bit of edge by manually increasing its estimation by some $\epsilon > 0$ at all time. Note that this procedure can change the gaps by at most ϵ, so our following discussions are not affected by this modification.

There are three quantities essential to the lower bound: 1). the term $\sum_{j=2}^{k} 1/\Delta_{[j]}^2$ which measures the *overall complexity* (total gap) of the problem instance; 2). the gap $1/\Delta_{[2]}^2$ between the best arm and the strongest contender; 3). the amount of variation (or divergence) in the gaps' distribution, as reflected by $H(\mathcal{I})$. In general,

the problem becomes more difficult when the gaps (especially $\Delta_{[2]}$) become smaller or more diverged.

The above result leads to our discussion on the challenges of adaptive online testing. Intuitively, adaptive exploration will be more efficient as we increase the number of arms. But now we understand how adding arms can also boost the complexity (either $\sum_{j=2}^{k} 1/\Delta_{[j]}^2$ or $1/\Delta_{[2]}^2$) or the entropy of the gaps. Consider a simple example with three arms and $\Delta_{[1]} = 1, \Delta_{[2]} = 1$. Suppose a new arm sits right on the middle of $a^{[1]}$ and $a^{[2]}$ and the gaps become: $\Delta_{[1]} = 1/2, \Delta_{[2]} = 1/2, \Delta_{[3]} = 1$. It means adding this new arm makes the problem four times more difficult. Non-adaptive online testing might just be as efficient in this case.

Not knowing the gaps in advance may not the the only factor that undermines adaptive online testing. Note that we previously assumed all arms have unit variance, which is unrealistic. Indeed, the variance associated with arm also plays a non-negligible role for the separability, and it actually acts as multiplicative factors on $\{1/\Delta_{[j]}^2\}_{j=2}^{k}$ [22]. Further, the confidence level will also become a stochastic process (such as the $\alpha^{(t)}$ in Section 3) instead of the fixed δ we just discussed in Lemma 3. The stochasticity in confidence level may further complicate the lower bound, and this scenario still requires new technical results to analyze. In light of these facts, conclusions cannot be made on how adaptive online testing is more effective than traditional A/B/n testing and why.

In conclusion, while recent advances from multiple domains have made adaptive online testing a practical and scientific alternative to traditional A/B/n testing, the inherent difficultly of the problem still poses challenges to fully embrace adaptive testing. Nevertheless, this technology has the potential to benefit multiple industries as more research and empirical work are underway.

5 FUTURE RESEARCH DIRECTIONS

We briefly discuss several promising directions that may contribute greatly to decision-making with adaptive online testing for IR.

- Extension to the fixed budget setting. The fixed budget setting may be more practical for certain IT productions such as advertising [30]. While the best-arm identification algorithms can often switch between the fixed confidence and fixed budget setting [14], the focus of hypothesis testing may vary. Increasing the power of a experiment design, rather than controlling the false discovery rate, might be more demanding for fixed budget problems. More research is needed to incorporate the new objective.

- *Thompson Sampling* (TS) for exploration. The Bayesian idea behind TS's exploration strategy is as powerful as the CB method [1], and Russo [29] recently developed TS-driven best-arm identification algorithm. A distinctive advantage of TS is to include prior knowledge about the experimentation (e.g. offline testing results). Although hypothesis testing is strictly a frequentist's choice, there is possibility to seek collaboration especially when TS involves only in exploration.

- Involving *contextual information*. Just like how contextual information serves multi-armed bandit, they can also be used to optimize adaptive online testing. For instance, when an online testing is designed to select the best banner size, the unknown rewards for each arm may be size-related. If so, an extra learning

step can lead to better exploration. Kato and Ariu [21] and Chen and Kasiviswanathan [10] has done some preliminary work in this direction, but this field remains largely unstudied.

- Real-world examination. Most of the methods in this short paper have only been tested via simulation or small-scale studies. Their efficacy in large-scale IT productions remains to be told. The real-time computation and updates required by adaptive online testing will certainly put a test on the infrastructure at the serving end. Business requirements may also complicate the design and analysis of adaptive online testing. Applied research with deployment solutions will be highly appreciated.

REFERENCES

[1] Shipra Agrawal and Navin Goyal. 2012. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*. JMLR Workshop and Conference Proceedings, 39–1.

[2] Ehud Aharoni and Saharon Rosset. 2014. Generalized α-investing: definitions, optimality results and application to public databases. *Journal of the Royal Statistical Society: Series B: Statistical Methodology* (2014), 771–794.

[3] Valentin Amrhein, Sander Greenland, and Blake McShane. 2019. Scientists rise up against statistical significance.

[4] Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. 2010. Best arm identification in multi-armed bandits.. In *COLT*. Citeseer, 41–53.

[5] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47, 2 (2002), 235–256.

[6] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 1 (1995), 289–300.

[7] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. 2013. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

[8] George Casella and Roger L Berger. 2021. *Statistical inference*. Cengage Learning.

[9] Lijie Chen, Jian Li, and Mingda Qiao. 2017. Towards instance optimal bounds for best arm identification. In *Conference on Learning Theory*. PMLR, 535–592.

[10] Shiyun Chen and Shiva Kasiviswanathan. 2020. Contextual online false discovery rate control. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 952–961.

[11] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. 2002. PAC bounds for multi-armed bandit and Markov decision processes. In *International Conference on Computational Learning Theory*. Springer, 255–270.

[12] Aleksander Fabijan, Pavel Dmitriev, Helena Holmstrom Olsson, and Jan Bosch. 2018. Online controlled experimentation at scale: an empirical survey on the current state of A/B testing. In *2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. IEEE, 68–72.

[13] Dean P Foster and Robert A Stine. 2008. α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70, 2 (2008), 429–444.

[14] Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. 2012. Best arm identification: A unified approach to fixed budget and fixed confidence. In *NIPS-Twenty-Sixth Annual Conference on Neural Information Processing Systems*.

[15] Aurélien Garivier and Emilie Kaufmann. 2016. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*. PMLR, 998–1027.

[16] Philip Hartman and Aurel Wintner. 1941. On the law of the iterated logarithm. *American Journal of Mathematics* 63, 1 (1941), 169–176.

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

[18] Adel Javanmard and Andrea Montanari. 2018. Online rules for control of false discovery rate and false discovery exceedance. *The Annals of statistics* 46, 2 (2018), 526–554.

[19] Ramesh Johari, Leo Pekelis, and David J Walsh. 2015. Always valid inference: Bringing sequential analysis to A/B testing. *arXiv preprint arXiv:1512.04922* (2015).

[20] Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. 2012. PAC subset selection in stochastic multi-armed bandits.. In *ICML*, Vol. 12. 655–662.

[21] Masahiro Kato and Kaito Ariu. 2021. The Role of Contextual Information in Best Arm Identification. *arXiv preprint arXiv:2106.14077* (2021).

[22] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. 2016. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research* 17, 1 (2016), 1–42.

[23] Emilie Kaufmann and Shivaram Kalyanakrishnan. 2013. Information complexity in bandit subset selection. In *Conference on Learning Theory*. PMLR, 228–251.

[24] Robert W Keener. 2010. *Theoretical statistics: Topics for a core course*. Springer Science & Business Media.

[25] Ron Kohavi and Roger Longbotham. 2017. Online Controlled Experiments and A/B Testing. *Encyclopedia of machine learning and data mining* 7, 8 (2017), 922–929.

[26] Tor Lattimore and Csaba Szepesvári. 2020. *Bandit algorithms*. Cambridge University Press.

[27] Edward Paulson. 1964. A sequential procedure for selecting the population with the largest mean from k normal populations. *The Annals of Mathematical Statistics* (1964), 174–180.

[28] Aaditya Ramdas, Fanny Yang, Martin J Wainwright, and Michael I Jordan. 2017. Online control of the false discovery rate with decaying memory. *Advances in neural information processing systems* 30 (2017), 5650–5659.

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

[30] Rajat Sen, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. 2017. Identifying best interventions through online importance sampling. In *International Conference on Machine Learning*. PMLR, 3057–3066.

[31] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[32] Fanny Yang, Aaditya Ramdas, Kevin Jamieson, and Martin Wainwright. 2017. A framework for multi-A (rmed)/B (andit) testing with online FDR control. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 5959–5968.

[33] Shengjia Zhao, Enze Zhou, Ashish Sabharwal, and Stefano Ermon. 2016. Adaptive concentration inequalities for sequential decision problems. *Advances in Neural Information Processing Systems* 29 (2016), 1343–1351.