Sequential Testing for Early Stopping of Online Experiments

Eugene Kharitonov[1,2], Aleksandr Vorobev[1], Craig Macdonald[2], Pavel Serdyukov[1], Iadh Ounis[2]

[1]Yandex, Russia
[2]University of Glasgow, UK

[1]{kharitonov, alvor88, pavser}@yandex-team.ru
[2]{craig.macdonald, iadh.ounis}@glasgow.ac.uk

ABSTRACT

Online evaluation methods, such as A/B and interleaving experiments, are widely used for search engine evaluation. Since they rely on noisy implicit user feedback, running each experiment takes a considerable time. Recently, the problem of reducing the duration of online experiments has received substantial attention from the research community. However, the possibility of using sequential statistical testing procedures for reducing the time required for the evaluation experiments remains less studied. Such sequential testing procedures allow an experiment to stop early, once the data collected is sufficient to make a conclusion. In this work, we study the usefulness of sequential testing procedures for both interleaving and A/B testing. We propose modified versions of the O'Brien & Fleming and MaxSPRT sequential tests that are applicable for testing in the interleaving scenario. Similarly, for A/B experiments, we assess the usefulness of the O'Brien & Fleming test, as well as that of our proposed MaxSPRT-based sequential testing procedure. In our experiments on datasets containing 115 interleaving and 41 A/B testing experiments, we observe that considerable reductions in the average experiment duration can be achieved by using our proposed tests. In particular, for A/B experiments, the average experiment durations can be reduced by up to 66% in comparison with a single step test procedure, and by up to 44% in comparison with the O'Brien & Fleming test. Similarly, a marked relative reduction of 63% in the duration of the interleaving experiments can be achieved.

Categories and Subject Descriptors: H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval

Keywords: A/B experiments; interleaving; statistical testing

1. INTRODUCTION

Online evaluation methods, such as A/B and interleaving experiments, have proven to be an important tool in web search evaluation [3, 9, 11]. In contrast to a Cranfield-based evaluation method, online evaluation methods interpret the implicit feedback of the real users. As a result, the online evaluation methods can be applied in several scenarios where offline evaluation might be impractical. Examples of such scenarios include cases where the large-scale labelling of the document relevance is hard, e.g. the evaluation of personalised ranking algorithms or the ranking of fresh content. Another example is the evaluation of changes in the search engine that do not affect search result ranking, such as changes in the UI, where A/B experiments can be applied.

However, the existing online methods have some limitations. As these methods rely on noisy implicit feedback from the users, each online evaluation experiment requires a considerable number of observations to be made before a reliable experiment outcome can be obtained [11]. As a result, usually each experiment is deployed on up to several per cent of query stream for several days, e.g. A/B experiments might be deployed for a week or two on several per cent of queries [11]. Similarly, Chapelle et al. [3] used a dataset of interleaving experiments that span up to 5 days. Due to these constraints, the number of simultaneously running experiments is bounded and, consequently, the throughput of the evaluation pipeline is limited. Since this limits the applicability of the online experimentation methods, an important research problem is to improve the online experimentation methods, so that they reach a reliable outcome faster.

Once the evaluated search engine modification is worse than the baseline system, the users who participate in the corresponding evaluation experiment are exposed to a somewhat degraded search experience. On the other hand, it was reported that such online experiments with negative outcome constitute a considerable part of all experiments [11]. Again, this supports the need to improve online evaluation so that the outcomes of experiments are obtained earlier: the negative changes are quickly rejected, and improvements are deployed as soon as possible.

These concerns become even more important in initiatives such as Living Labs [1], where industrial participants are invited to share a part of their live traffic for use by academic researchers. Indeed, the ability to quickly detect and disable an erroneous experiment can be extremely useful in such a shared infrastructure.

As we will discuss in Section 2, several approaches to reduce the duration of interleaving experiments have previously been studied. However, the possibility to reduce the duration of the interleaving experiments by using *sequential testing procedures* remains less studied. Such testing procedures are capable of stopping the experiments early once the available data is sufficient to make a reliable comparison outcome. In this work, we aim to close this gap. We demonstrate that by modifying two sequential testing procedures to make them applicable for interleaving, we obtain a considerable improvement in the average time an experiment takes.

Similarly, to the best of our knowledge, there are no studies published that quantitatively assess the usefulness of sequential testing in the context of the A/B experiment-based search evaluation. In this work, we propose a new sequential test developed to match the distribution of one of the popular absolute online metrics, namely the abandonment rate. Moreover, we perform an experimental study on a dataset of A/B experiments, where we evaluate both our proposed test and the standard O'Brien&Fleming sequential test.

Overall, the contributions of this work are two-fold:

- We propose several sequential testing methods that reflect the distributions of the data generated in A/B and interleaving experiments, and describe how to adjust their stopping thresholds based on query log data;

- We perform an extensive evaluation study of the performance of our proposed methods using real-life datasets of A/B and interleaving experiments.

The remainder of this paper is organised as follows. In Section 2 we discuss the related work. In Section 3 we briefly review how A/B test and interleaving experiments are performed. In Section 4 we introduce several methods for performing sequential statistical analysis. The datasets used in our empirical study are described in Section 5. Our evaluation methodology and the results we obtained are described in Sections 6 and 7, respectively. We conclude this paper and discuss future work in Section 8.

2. RELATED WORK

Our work is closely related to two areas of research. The first area is concentrated on developing approaches to speed up the existing online evaluation methods and we discuss these approaches in Section 2.1. The sequential analysis methods are reviewed in Section 2.2.

2.1 Improving online evaluation

The first interleaving method, Balanced Interleaving, was proposed by Joachims [7]. The method was further developed in Team Draft Interleaving [18], Probabilistic Interleaving [6], Optimised Interleaving [17], and their modifications.

A thoroughly studied approach to reduce the time required to run interleaving experiments is to reduce noise in the user feedback and thus to improve the convergence rate of the interleaving experiments. This can be achieved by modifying the way the user clicks are aggregated and the aggregated credit is assigned to the tested alternatives. Radlinski and Craswell [16] studied approaches to weight user clicks according to the ranks of the clicked results. Chapelle et al. [3] considered ten different heuristic click aggregation schemes. For instance, in one of the schemes the result of each interleaved impression is represented by the difference between clicks obtained by compared systems, divided by the total number of clicks in the corresponding impression. In another scheme, the result is binary, equal to -1 if the baseline system received more clicks than the tested, +1 in the opposite case, and equal to 0 if they receive equal number of clicks. Overall, Chapelle et al. found that some of the tested credit assignment rules improve convergence rate of the interleaving outcomes.

A more elaborated approach to interpret click feedback was considered by Yue et al. [24], who proposed to machine-learn the click aggregation scheme using a feature-based representation of the clicks. Under their proposed approach, the combination of the features is learned to maximise the confidence in the outcomes of the earlier performed experiments.

Another approach was proposed by Radlinski and Craswell in [17]. Based on a novel interleaving framework, they studied how a sensitive interleaving method (i.e. the method that quickly converges) can be built by means of maximising the uncertainty of the comparison winner within a single session. Building on their work, Kharitonov et al. [8] proposed an approach to leverage pre-experimental data to improve the interleaving sensitivity.

Deng et al. [4] studied an approach to improve the convergence rate of A/B experiments by using pre-experimental data. This reduction is achieved, for instance, by means of stratification of the users into groups, such that the between-strata variance of the observed metric is removed.

All of the above discussed approaches share the same statistical testing scenario. In this scenario, an online evaluation experiment is deployed. After running this experiment for a pre-defined period of time (e.g. a week), the experiment is stopped. Next, some form of statistical test, such as the binomial test in the case of interleaving, is performed on the collected user interaction data to infer if a statistically significant difference between the tested alternatives was observed. Importantly, in this scenario, each experiment is deployed for the period of time that is fixed before the experiment starts. However, it is likely that in some experiments highly contrasting alternatives are compared, and in that case it should be possible to stop the experiment early and still be able to reliably detect preferences between the compared alternatives. In this paper, we study how to conduct a statistical analysis which is capable of stopping such experiments early, and thus reduces the mean experiment evaluation time. In our experimental study, we demonstrate that our proposed early-stopping approach is complementary to the approaches studied before, and the online evaluation methods can benefit from combining approaches to improve sensitivity and to stop early.

2.2 Sequential testing

Sequential statistical testing appeared to address the demands of the military testing during World War II, resulting in Wald's *sequential probability ratio test* (SPRT) [20]. This test performs an analysis in steps. At each step, a new data point is considered, and the decision is taken if the observed data is enough to make a reliable conclusion about the considered hypotheses and the experiment should be finished, or more measurements are required.

Despite its simplicity, SPRT was shown [21] to be optimal when comparing two simple alternatives. Further research was conducted to improve the SPRT-based methods in a variety of directions. For instance, the 2-SPRT test was proposed to minimise the expected sample size at a specified parameter when discriminating hypotheses $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1 > \theta_0$ [2]. Another SPRT-based test that can be used to test against a complex hypothesis, MaxSPRT [12], was proposed for the post-approval drug safety surveillance. These tests can be applied when the specific parameters of one of the tested hypothesis are not known before running the experiment. A similar problem arises while running online experiments. Indeed, in interleaving experiments the null hypothesis can be specified (both evaluated algorithms are equally likely to win in a particular user session), but the difference between the evaluated algorithms is hard to estimate before running an experiment.

As an alternative to the SPRT-based tests, "repeated significance tests" (RST) were proposed. As conventional single-sample tests are often used in clinical trials, the motivation behind RST is to apply them repeatedly during the trial. These tests evolved into a group sequential RST [15], where the data is accumulated between tests. Further, O'Brien and Fleming [14] proposed a group sequential testing procedure that had a better performance. These methods became popular within clinical trials, as the sequential procedures reduce the time the participants are exposed to ineffective or harmful treatments.

Thereafter, the group sequential testing approach was intensively developed. Wang and Tsiatis [23] suggested a parametric family of tests, which generalises both Pocock and O'Brien&Fleming tests, and can be optimised to be nearly optimal w.r.t. a fixed expected difference between the alternatives. Another improvement was proposed by Lan and DeMets [13], which accounts for some specifics of the clinical trials: the number of subjects available between stops is not known in advance and can vary greatly.

Overall, sequential testing is a highly developed discipline, and a variety of tests that differ by their properties and assumptions was proposed. A review can be found in [2, 19]. Due to their properties, we select the O'Brien&Fleming and MaxSPRT tests as a foundation for our study. These tests are very practical, they do not require any a priori assumptions about the expected effect size or its boundaries. The O'Brien&Fleming test can be interpreted as a repeated standard Pearson's chi-square test with progressive stopping thresholds, so that the last threshold is close to the one test scenario, which is very appealing from the practical perspective. Similarly, the MaxSPRT-based tests do not require any pre-experimental knowledge, and their decisions are extremely transparent.

Finally, to the best of our knowledge, the only work that mentions the use of sequential testing procedures in online web search evaluation is [11]. In their work, Kohavi et al. reports that the O'Brien&Fleming test is used in Bing to abort A/B experiments early when a severe degradation in metrics is observed. In contrast, we propose modifications of the MaxSPRT tests for interleaving and A/B experiments, and modify the O'Brien&Fleming test for the interleaving evaluation. Moreover, we perform a thorough evaluation of the usefulness of the considered tests.

3. ONLINE EXPERIMENTATION

In this section we briefly review how the A/B and interleaving experiments are performed.

3.1 A/B tests

To perform an A/B comparison of a modified system B to the baseline system A, each of the systems is assigned to serve its own bucket of users (a bucket is a random sample of users). Usually, each experiment lasts for a week or two [11], so that the data collected contains sessions from each day of the week. After stopping the experiment, some absolute online metrics are measured on each bucket. These metrics are compared across buckets, and the statistical significance of the difference of the metrics is analysed, e.g. by means of a t-test. More formally, denoting the expected values of the metric μ on buckets A and B, as μ_A and μ_B respectively, we are interested in comparing two statistical hypotheses, H_0 and H_1:

$$H_0 : \mu_A = \mu_B, \quad H_1 : \mu_A \neq \mu_B \tag{1}$$

A variety of online absolute metrics were proposed, and, usually, a number of metrics is used in combination. For instance, when optimising for revenue one needs to ensure that there is no degradation in user satisfaction metrics [10]. Additionally, a set of diagnostic metrics might be used to automatically terminate a deployed experiment, when a severe degradation in one of the metrics is observed [11]. The early stopping procedures we study in this work are especially useful in controlling the changes in the diagnostic metrics.

In our study we use the abandonment rate metric, which is equal to the fraction of the user interactions with no results clicked on. We believe that the extension of our study to account for other metrics such as those representing user engagement [5] is a promising direction for future work.

3.2 Interleaving

All of the existing interleaving methods share the same idea. Suppose we want to compare two alternative ranking algorithms, A (baseline) and B (evaluated system). To compare them, we randomly select a sample of user sessions in the session stream and use them in the experiment. For each query submitted in the experiment, the results from A and from B are retrieved. In the next step, these results are mixed (interleaved) and then presented to the user.

While the approaches we discuss in our work can be applied for any existing interleaving methods, below we review the Team Draft [18] interleaving method, as it is used in our evaluation study. The interleaved result list is generated from the result lists of A and B by the following algorithm. The first result is selected from a random alternative. After that, from the second alternative, we take the first result that is not already included in the interleaved result list. These two steps are repeated until the required number of results is obtained.

As we discussed in Section 2, several approaches to aggregate the user clicks observed in a session into a credit obtained by A and B were proposed [3, 16, 24]. In this study, we experiment with two aggregation schemes: the *binary*, and the *deduped binary* schemes. Under the binary scheme, in each interaction, the alternative with the most results clicked receives a unit credit and is referred to as the winner. The deduped binary scheme is similar, but clicks on the top-k results which are identical both in A and B are ignored. In both schemes, interactions without clicks are ignored. If in a session both A and B obtained an equal number of clicks, the session is considered as tie.

The deduped credit assignment removes some additive zero-mean noise from the user feedback. Despite its simplicity, it was reported [3] that the deduped binary scheme is one of the top-performing schemes, markedly reducing the number of user interactions required to obtain a statistically significant experiment outcome.

Let us denote a variable S that represents the probability of B winning A in a session with a click, assuming that the ties are broken randomly. After running an experiment, S can be estimated as follows:

$$\hat{S} = \frac{w^B + \frac{1}{2}t}{w^A + t + w^B} \tag{2}$$

where w^B and w^A denote the number of sessions where B and A win, respectively; t is the number of ties.

The goal of the statistical analysis methods we discuss in Section 4 is to compare two statistical hypotheses, H_0 (A and B are equally likely to win a particular impression) and H_1 (the chances to win differ):

$$H_0 : S = \frac{1}{2}, \quad H_1 : S \neq \frac{1}{2} \tag{3}$$

4. STATISTICAL ANALYSIS

In this section, we introduce the sequential testing procedures we consider in this work. We start by describing the procedures applicable for interleaving experiments (Section 4.1): O'Brien&Fleming's sequential test, modified for interleaving (OBF-I), and the MaxSPRT test. After that, we describe two tests applicable for A/B tests: the standard O'Brien&Fleming test, and our proposed MaxSPRT-AB test, tailored for A/B test experiments. For all of these tests, we also describe algorithms to train their stopping thresholds.

4.1 Interleaving

While analysing an interleaving experiment, our goal is to compare two statistical hypotheses (Equation (3)): under the null hypothesis (H_0), A and B have equal chances to win an interleaving comparison in an interaction; under the alternative hypothesis (H_1) these chances are not equal.

OBF-I Initially, the O'Brien & Fleming's sequential test was formulated for clinical trials that compare two treatments on two different groups of participants. In contrast, in interleaving experiments only one group of users is used. Below we describe our adaptation OBF-I of the OBF test to the case of interleaving experiments.

Assume that the number of possible stops, where a sequential test is allowed to analyse accumulated data and make a decision, is set to N. Let us introduce a random variable x that is equal to 1 (-1) if B (A) wins in a comparison in an interaction, and 0 if a tie is observed. By x_j we denote the realisation of x observed in jth session. Further, we denote the number of sessions between the $(i-1)$th and ith stops as K_i. Under the null hypothesis, the probabilities of winning a comparison in a session for A and B are equal. Thus, according to the central limit theorem, the normalised mean R_i of the realisations x_j observed between the $(i-1)$th and ith stops approaches the standard normal distribution as K_i grows:

by stops we can, for example, denote days

$$R_i = \frac{(x_1 + .. + x_{K_i})}{(K_i \cdot D[x])^{\frac{1}{2}}} \sim \mathbb{N}(0, 1) \tag{4}$$

where $D[x]$ is an estimate of the variance of x.

Further, we denote the total number of sessions occurred before the ith stop as T_i ($T_i = \sum_{j<i} K_j$), and the accumulated number of comparisons won by A (B) before the ith stop as w_i^A (w_i^B). Assuming that the number of the sessions occurring between the stops is approximately the same and equal to K, we define the accumulated statistic $O_i = (\frac{1}{\sqrt{i}} \sum_1^i R_j)^2$. Since $\sum_1^i R_j$ is a sum of variables that are distributed according to $\mathbb{N}(0, 1)$, their scaled sum $\frac{1}{\sqrt{i}} \sum_1^i R_j$ also has the standard normal distribution. Thus, as a square of a standard normal variable, O_i is distributed according to the chi-squared distribution with one degree of freedom:

$$O_i = \left(\frac{1}{\sqrt{i}} \sum_{j=1}^{i} R_j \right)^2 = \frac{(w_i^B - w_i^A)^2}{iK \cdot D[x]} = \frac{(w_i^B - w_i^A)^2}{T_i \cdot D[x]} \sim \chi^2(1) \tag{5}$$

The estimate of the variance $D[x]$ is:

$$D[x] = \frac{1}{T_i - 1} \sum_{j=1}^{T_i} (x_j - \bar{x})^2, \quad \bar{x} = \frac{w_i^B - w_i^A}{T_i} \tag{6}$$

O'Brien and Fleming [14] proposed to apply a progressive decision criterion where at the ith stop, O_i is compared to a threshold $\frac{1}{i}\hat{a}$ that decreases at each stop ($\hat{a}$ depends on the number of stops and required Type I error). This ensures an intuitive requirement that to terminate an experiment earlier, one needs to have a higher confidence in H_1.

In an equivalent but more convenient formulation, at each stop, a statistic $i \cdot O_i$ can be considered, and compared to a single fixed threshold a. Once it exceeds a, the experiment is terminated, and H_1 is accepted. To infer the experiment outcome, the difference between w_i^A and w_i^B is used (i.e. if $w_i^B > w_i^A$ then $B \succ A$). If at the last stop $N \cdot O_N$ still does not reach a then the hypothesis H_0 is accepted.

For the cases of small numbers of stops (less or equal to 5), the values of the threshold a can be found in [14]. Since in our experiments we use a higher number of stops, we briefly review how the threshold can be obtained from running Monte-Carlo simulations. The general idea is to replace R_i with random numbers generated from $\mathbb{N}(0, 1)$, and adjust a so that the test will detect a difference in the α (required Type I error level) fraction of the generated tests. Formally, to perform one iteration of the simulation, we sample N (N is the required number of stops) random numbers from the standard normal distribution ($U_1, ..., U_N \sim \mathbb{N}(0, 1)$), and calculate the maximum square of their partial sums $U_m^2 = \max(U_1^2, (U_1+U_2)^2, ..., (U_1+...U_N)^2)$. We collect these maximums over 10,000 simulations. Finally, we select a value that corresponds to $(1 - \alpha)$ percentile of these maximums.

A possible heuristic is to replace $D[x]$ with its upper bound[1] 1. While this substitution might increase the time required for O_i to achieve the threshold a, it also makes the decision rule even simpler: at each stop i, a normalised square of the difference between the wins of A and B is multiplied by the number of the stop and compared to the threshold a. We refer to a rule with this heuristic applied as **OBF-I***.

Notably, OBF-I and OBF-I* assume that the number of sessions performed between stops is large enough so that the central limit theorem can be applied.

MaxSPRT At the core of the SPRT family of tests is the likelihood ratio statistic. Informally, this statistic equates to the likelihood of the observed data under the alternative hypothesis H_1 divided by the likelihood of the data under H_0. Once this ratio becomes big enough, H_0 can be rejected. To formalise this idea, we use the same notation as before. By T_i we denote the total number of sessions before the ith stop, w_i^A (w_i^B), and t_i are the numbers of sessions where A (B) wins, and the number of sessions with ties, respectively. Further, by m_i we denote our estimate of the number of the comparisons won by B after breaking the ties:

$$m_i = w_i^B + \frac{1}{2} t_i$$

Under this notation, the logarithm of the likelihood statistic can be specified as follows:

$$L_i = \log \frac{p_1^{m_i} (1 - p_1)^{T_i - m_i}}{p_0^{m_i} (1 - p_0)^{T_i - m_i}} \tag{7}$$

[1] A unit variance is achieved if on each tie a coin is tossed and a unit credit is assigned to a random alternative. However, this might be a good approximation since for a tie to occur at least two results must be clicked, which happens rarely.

$$\frac{\bar{X} - \mu}{3/\sqrt{n}} \to N(0,1) \quad , \quad X_i \in \{-1, 0, 1\}$$

under H_0 : $E\bar{X} = 0 \wedge \mu = 0$

$$\Rightarrow \frac{\sum X_i}{n(3/\sqrt{n})} = \frac{\sum X_i}{3\sqrt{n}}$$

```
Input: Type I error tolerance $\alpha$, a set of A/A
       experiments $Q$.
Output: $\bar{L}$ threshold.
//the vector of the ratio values observed in experiments
$Ls \leftarrow \emptyset$
//iterate over experiments
foreach $e \in Q$ do
    //iterate over sessions in $e$
    $L_m \leftarrow 0$
    foreach $i \in 1..|e|$ do
        $T_i \leftarrow w_i^B + t_i + w_i^A, \quad m_i \leftarrow w_i^B + \frac{1}{2}t_i$
        //find the max. likelihood estimate $\hat{p}_1^i$ of $p_1$
        $\hat{p}_1^i \leftarrow \frac{1}{T_i}\left[w_i^B + \frac{1}{2}t_i\right]$
        $L_i \leftarrow \log \frac{(\hat{p}_1^i)^{m_i}(1-\hat{p}_1^i)^{T_i - m_i}}{p_0^{m_i}(1-p_0)^{T_i - m_i}}$
        //update the maximum value of $L_m$ for the
        current experiment
        $L_m \leftarrow \max(L_m, L_i)$
    end
    $Ls \leftarrow Ls \bigcup \{L_m\}$
end
$Ls \leftarrow sorted(Ls)$
$\bar{L} \leftarrow Ls[|Ls| \cdot (1-\alpha)]$
```

Algorithm 1: Learning the $\bar{L}$ threshold for MaxSPRT from a dataset of A/A experiments.

where p_0 and p_1 are probabilities of B winning in a comparison in an interaction under H_0 and H_1, respectively. Under the null hypothesis, the alternatives are equally likely to win, so p_0 equals to $\frac{1}{2}$. However, it is hard to specify p_1 before actually running the experiment. An intuitive idea is to replace it with the maximum likelihood estimate $\hat{p}_1^i$, based on the experimental data observed before the ith stop. Informally, by estimating $\hat{p}_1^i$ we choose H_1 that is the most likely to be accepted in comparison with H_0. This idea was proposed and studied by Kulldorff et al. [12] for the Poisson and Binomial distributions, and resulted in a test called MaxSPRT. Under our notation, the maximum likelihood estimate of p_1 at the ith step is:

$$\hat{p}_1^i = \frac{1}{T_i}\left(w_i^B + \frac{1}{2}t_i\right)$$

At each stop, $\hat{p}_1^i$ is estimated, and is used as a substitute of p_1 in L_i (Equation (7)). After that, L_i is compared to a pre-defined threshold $\bar{L}$. If $L_i \geq \bar{L}$, then the experiment is stopped, and H_1 is accepted. If $\hat{p}_1^i > \frac{1}{2}$ ($\hat{p}_1^i < \frac{1}{2}$) then it is inferred that $B \succ A$ ($A \succ B$). If $L_i < \bar{L}$, the experiment is continued. H_0 is accepted if the experiment reaches a pre-defined maximum length, without achieving $\bar{L}$.

To specify the threshold $\bar{L}$, Kulldorff et al. [12] used a Monte-Carlo method, where a series of Binomial samples were generated. However, as we further discuss in Section 7, in the case of the interleaving experiments where the ties are interpreted according to Equation (2), this Monte-Carlo threshold adjustment is suboptimal, as it generates data with variance higher than observed in experiments.

Instead, we propose to train the threshold $\bar{L}$ on a set of experiments where a system is compared with itself. As we discuss further in Section 5, such experiments are referred to as A/A experiments. Intuitively, a statistical test with Type I error set to α should detect differences in A/A experiments approximately with the probability α. Using this idea, we adjust the threshold $\bar{L}$ so that $\bar{L}$ exceeds all values of L_i in $(1-\alpha)$ of the A/A experiments. A formal description of the optimisation of the threshold can be found in Algorithm 1.

Further, the MaxSPRT test with the threshold $\bar{L}$ trained using Monte-Carlo simulations is denoted as **MaxSPRT-I-MC**. The test with the threshold $\bar{L}$ trained using the A/A comparisons is referred to as **MaxSPRT-I-AA**.

Note that when the deduped binary click scoring scheme is applied, all clicks in A/A experiments are ignored, since in the results lists of the compared alternatives all their results are identical. However, we still want our test not to detect a difference between systems when the results are not identical documents, but are equally likely to be clicked, to satisfy the user. Thus, we use the thresholds trained using the binary click aggregation scheme when evaluating the test on the experiments with the binary deduped scheme applied.

The MaxSPRT test assumes that the data points arrive one-by-one, which might be impractical on modern large-scale web search engines. Indeed, an infrastructure is needed that is capable of providing a near real-time stream of individual comparisons. It can be easier to implement the data delivery in batches, e.g. each batch of data corresponding to an hour or a day of the user activity. Since the discussed tests can be applied to analyse batch data by simply considering the aggregated values of the variables such as w_i^A w_i^B, we experiment in the batch scenario.

4.2 A/B tests

The goal of the statistical analysis of the A/B tests is to compare two hypotheses (Equation (1)). Under the null hypothesis, the means of a considered metric are equal in the buckets assigned for A and B, while under the alternative hypothesis they are different.

OBF The original test proposed by O'Brien&Fleming can be directly applied to A/B tests. Below we shortly review how it is defined. By $\hat{\mu}_i^A$ and $\hat{\mu}_i^B$ we denote the sample mean values of a metric μ calculated using the data collected before the ith stop on buckets (samples of users) associated with A and B, respectively. The total number of sessions that took place before the ith stop are referred to as T_i^A and T_i^B. The number of stops is fixed to N, the number of sessions between stops is assumed to be roughly equal and large enough for the central limit theorem to be applicable.

At each stop, the following statistic is considered:

$$Z_i = i\frac{\left(\hat{\mu}_i^A - \hat{\mu}_i^B\right)^2}{D[\hat{\mu}^A - \hat{\mu}^B]} \qquad (8)$$

We follow [14] and use the pooled estimate of variance $D[\hat{\mu}^A - \hat{\mu}^B]$ assuming H_0:

$$\begin{aligned} D[\hat{\mu}_i^A - \hat{\mu}_i^B] &= \left(\tfrac{1}{T_i^A} + \tfrac{1}{T_i^B}\right) D[x], \\ D[x] &= \tfrac{1}{T_i^A + T_i^B - 1}\left(\sum_{j=1}^{T_i^A + T_i^B}(x_j - \overline{x})^2\right) \end{aligned} \qquad (9)$$

where x_j iterates over the metric values observed in all observed interactions before the ith stop.

Similarly to the interleaving case, at each stop the value of the statistic Z_i is compared to a constant threshold a. Once it reaches the threshold, the experiment is stopped and H_1 is accepted. If at the last stop Z_N is still less than a, H_0 is accepted. The threshold a is calculated using the same Monte-Carlo algorithm as in the case of interleaving.

MaxSPRT-AB The MaxSPRT rule cannot be easily applied for the case of A/B experiments, since the distribution of the considered metric is unknown both under H_0 and H_1. To address this, we propose to estimate both the distribution of the metric under the null hypothesis and under the alternative hypothesis using data from the experiment itself.

Under the null hypothesis, the distribution of the metric coincides in both buckets, but under H_1 the distributions are different. Further, by D_i^A (D_i^B) we denote the data observed in a bucket corresponding to A (B) upto the ith stop. In that case, the likelihood ratio statistic in the case of the abandonment rate metric can be represented as follows:

$$L_i = \log \frac{P(D_i^A, D_i^B | p_i^A, p_i^B)}{P(D_i^A, D_i^B | p_i^0)} \tag{10}$$

where p_i^A and p_i^B are the maximum likelihood estimates of the probabilities of a user abandoning the result page, obtained on buckets corresponding to A and B, respectively. Similarly, p_i^0 is the maximum likelihood estimate of the probabilities of a user abandoning the result page, calculated on both buckets. $P(D_i^A, D_i^B | p_i^A, p_i^B)$ is calculated according to Equation (11):

$$\begin{aligned} P(D_i^A, D_i^B | p_i^A, p_i^B) = \quad & (p_i^A)^{C_i^A} (1 - p_i^A)^{T_i^A - C_i^A} \cdot \\ & \cdot (p_i^B)^{C_i^B} (1 - p_i^B)^{T_i^B - C_i^B} \end{aligned} \tag{11}$$

where C_i^A and C_i^B are the number of abandoned result pages for A and B buckets before the ith stop, respectively. Similarly, T_i^B and T_i^A are the total numbers of sessions in these buckets before the ith stop.

Under the null hypothesis, the likelihood of the observed data is calculated as follows:

$$\begin{aligned} P(D_i^A, D_i^B | p_i^0) = \quad & (p_i^0)^{C_i^A} (1 - p_i^0)^{T_i^A - C_i^A} \cdot \\ & \cdot (p_i^0)^{C_i^B} (1 - p_i^0)^{T_i^B - C_i^B} \end{aligned} \tag{12}$$

The maximum likelihood estimates p_i^0, p_i^A, and p_i^B can be found as follows:

$$p_i^0 = \frac{C_i^A + C_i^B}{T_i^A + T_i^B}, \quad p_i^A = \frac{C_i^A}{T_i^A}, \quad p_i^B = \frac{C_i^B}{T_i^B} \tag{13}$$

Once L_i achieves a pre-fixed threshold $\bar{L}$, the experiment is stopped and H_1 is accepted. Otherwise, the experiment is continued. If the experiment reaches a fixed duration, H_0 is accepted. Again, the threshold is learned by finding the $(1 - \alpha)$ percentile of the distribution of L_i statistic on the dataset of A/A experiments, as in Algorithm 1.

5. DATASETS

In our evaluation study we use datasets of A/B and interleaving experiments obtained from Yandex. For diagnostic purposes, it is useful for a search engine to deploy a constantly running online experiment that compares the current production system with itself ([10], Section 2). Further, we refer to such an experiment as an A/A comparison, whether the comparison is performed by A/B test or by means of interleaving. Since we know that the alternatives are equal in this comparison, we want the statistical testing procedure to find statistical differences in this evaluation rarely (i.e. H_0 should be rejected about 1% of the time, when testing is performed on $p < 0.01$ significance level).

Another source of the experiments are the regular experiments that are deployed to evaluate new search engine improvements. In our evaluation study, we compare the sequential testing approach to a standard scenario, where experiments are deployed for an integer number of weeks. To increase the size of the dataset, we consider the case of the experiments that last for one week. However, as the experiments differ in the expected effect size (detecting smaller differences between A and B require more sessions), they also vary in their duration. For this reason, we restrict each experiment to its first 7 days. The ground-truth outcomes used in our evaluation are calculated using the full experimental data. However, the one-step baseline tests (t-test and binomial test) use the same restricted experimental data as is provided for the evaluated sequential tests.

Interleaving In our dataset of interleaving experiments, we include data from two interleaving-based A/A experiments over a period of 300 days during 2014. These two experiments were deployed in two different countries. Further, we sampled 206 real-life interleaving experiments that were deployed to evaluate changes in the ranking algorithms, and lasted for at least a week during the same 300 days. Among these experiments, a statistically significant outcome (binary credit scheme, $p < 0.001$, binomial test) is observed in 115 experiments (B outperforms A in 56 experiments).

A/B tests Similarly to the interleaving dataset, we included the query log data generated by two long-term A/A tests that compare the production system with itself. This data spans a two month period, from April to May, 2014. These two experiments were deployed in two countries. Further, we also sample 62 A/B tests that were performed as a part of the search engine's evaluation routine. From these experiments, we select 41 experiments that have a statistically significant outcome (abandonment rate metric, $p < 0.01$, t-test). In 14 experiments, B outperformed A.

6. EVALUATION METHODOLOGY

We split our discussion of the evaluation methodology in two sections. In Section 6.1 we introduce the quality metrics we use in our evaluation. In Section 6.2 we describe the evaluation procedure we use.

6.1 Metrics

Our first metric, **Type I error**, represents the probability of a statistical test rejecting the null hypothesis H_0 when it holds:

$$\alpha = P(H_1 \ accepted \ | H_0 \ holds)$$

Generally, we want Type I error to be low, as each experiment wrongly accepted as successful might result in expensive development, wastes both human and computational resources without improving the search engine.

The second metric we consider is **Type II error**, which measures the probability of accepting the null hypothesis when it does not hold:

$$\beta = P(H_0 \ accepted \ | H_1 \ holds)$$

High values of β indicate that non-equal alternatives A and B are frequently accepted as equal, and this results in ignoring opportunities to improve a search engine.

The Type II error metric defined as above does not penalise cases when the null hypothesis is correctly rejected, but the preference is inferred incorrectly (e.g. $A \succ B$ is accepted when in reality $B \succ A$)[2]. Thus we introduce two one-sided accuracy metrics, $Acc_{A \succ B}$ and $Acc_{B \succ A}$:

$$\begin{aligned} Acc_{A \succ B} &= P(accepted \ that \ A \succ B | A \succ B) \\ Acc_{B \succ A} &= P(accepted \ that \ B \succ A | B \succ A) \end{aligned}$$

These metrics are related to Type II error, however they additionally penalise the above discussed cases of the incorrectly inferred preferences.

[2] This situation arises as our tests have effectively three outcomes: $A \succ B$, $B \succ A$, and B is not different from A.

Mean deployment time, $\mathbb{E}(T)$. This metric is defined as the mean time the experiment is deployed before a sequential testing procedure stops. For the non-sequential one-step tests that we use (namely, t-test and binomial sign test), the value of this metric is set to the experiment length. For convenience, we measure this metric in days. Generally, it might be more important to stop an experiment where $A \succ B$ than an experiment where $B \succ A$, as in the former case the user experience is degraded. Thus we additionally consider two time-related metrics which measure the expected duration of the experiments where $A \succ B$ and $B \succ A$. We denote these metrics as $\mathbb{E}(T|A \succ B)$ and $\mathbb{E}(T|B \succ A)$.

Mean relative number of sessions, $\mathbb{E}(\frac{N}{N_0})$. The last metric we use represents the relative number of search interactions required until the experiment is stopped, averaged over all experiments in the dataset. In other words, if an experiment contains N_0 sessions (after restricting to its first 7 days), and the sequential testing procedure stops an experiment after observing only N sessions, the value of the metric is equal to $\frac{N}{N_0}$ on this experiment.

In our evaluation study, for each sequential testing procedure we fix the Type I error probability and the maximum deployment time to be the same. Under these constraints, the baseline one-step approach achieves the minimum Type II error level among all possible rules. Thus, *our goal is to find a sequential testing procedure that reduces the mean deployment time for the experiments in our dataset as much as possible, and has its Type II error level close to the baseline approach.*

6.2 Evaluation Protocol

The A/A experiments, which compare a system with itself, either by means of interleaving or in a A/B test, are a perfect source of the data to calculate the Type I error probability (i.e. probability of finding a difference when there is none). Indeed, in the A/A experiments the null hypothesis H_0 definitely holds. Thus, any event where a statistical test detects a difference between the tested alternatives in an A/A experiment, is a Type I error.

Using this observation, we apply the following scheme to measure the Type I error probability for a statistical test. First, we split each of the available A/A experiments in two non-overlapping parts. The first part is used for learning the stopping thresholds. The second part is used for calculating the Type I error probability. To calculate this probability, we generate a set of smaller experiments of length equal to the length of the real experiments in the dataset (7 days). This set is generated by moving a sliding window from the beginning of the A/A experiment towards its end. The length of the window is equal to 7 days for both the A/B and interleaving experiments. We measure the Type I error probability as the relative number of generated experiments where the testing procedure detects a difference between two compared systems. The initial splitting is repeated in the cross-validation process, so that each day of the experiment is included in the evaluation subset exactly once. We use 20-fold cross-validation in the case of interleaving, and 10-fold for A/B tests (due to a smaller size of the dataset).

In the evaluated tests, we set the tolerances for the Type I error to be 0.01 and 0.05 for the interleaving and the A/B experiments, respectively. Generally, we want to experiment with lower tolerance values, as this closer resembles the requirements for real experiments. These tolerance levels should be higher than the p-values we use to infer the ground-truth labels, so that the measurements are meaningful. In turn, using low p-values when obtaining the ground-truth labels significantly reduces the sizes of the datasets. Thus we believe that the selected values are reasonable.

To calculate the remaining metrics (Type II error, $Acc_{A \succ B}$, $Acc_{B \succ A}$, $\mathbb{E}(T)$, $\mathbb{E}(\frac{N}{N_0})$), we use the experiments that compare real-life changes in a search engine. As a ground-truth labels ($A \succ B$ or $B \succ A$) we use the results of the binomial test ($p < 10^{-3}$) for the interleaving experiments, and the t-test ($p < 10^{-2}$) for A/B experiments.

An alternative approach to calculate Type II errors is to use a set of experiments, where the tested alternative B is specifically degraded with respect to A. This degradation might be achieved by swapping the first and the second results, degraded snippets, using an inferior ranking algorithm, etc. However, building a big dataset of experiments, where the user experience is specifically degraded in different ways, can be unrealistic. Another concern is that such manually devised degradations cannot be considered as a representative sample of the real-life experiments, and thus measuring how they can be accelerated might be useless.

In our evaluation study, we vary the number of the stops used. In the first case, the stops are performed each day (i.e. 7 stops), and in the second case stops are performed every hour (i.e. $7 \cdot 24 = 168$ stops). While the SPRT-based tests can be applied on the per-interaction level, the gains from such a scenario cannot be more than an hour in comparison with the scenario with stops every hour.[3] On the other hand, to apply a per-interaction SPRT test, one would need to build an elaborated near real-time data delivery system.

7. RESULTS AND DISCUSSION

In this section we discuss the results of our evaluation study on the datasets of interleaving (Section 7.1) and A/B (Section 7.2) experiments. Further, we perform a visualisation of the decisions of the best-performing tests in Section 7.3.

7.1 Interleaving

By OBF-I we denote the adaptation of the O'Brien&Fleming test which we discussed in Section 4.1. By OBF-I* we denote the simplified modification of OBF-I that approximates the variance by the unity. MaxSPRT-I-MC is the MaxSPRT test with its $\bar{L}$ threshold trained by the Monte-Carlo approach. In contrast, MaxSPRT-I-AA corresponds to the MaxSPRT test with its $\bar{L}$ threshold trained on the dataset of A/A experiments by Algorithm 1.

In Tables 1 & 2 we report the results of the evaluation of the sequential testing rules on the dataset of the interleaving experiments, in the cases of the binary and deduped binary click aggregation schemes, respectively. On analysing these results we firstly notice that the Type I error levels measured for all the considered testing rules are close to the tolerance level we set in the threshold learning process, 0.01. The observed deviations might be caused by a limited size of the dataset we use.

Further, on analysing the Type II error metric reported in Table 1, we notice that the values of this metric are very close for all tests and are in the range of $0.10 - 0.13$, except for the MaxSPRT-I-MC test, which has higher error levels (e.g. Type II error probabilities are 0.23 and 0.19, for the cases with 7 and $7 \cdot 24$ stops, respectively). In the case

[3] Due to the design of the test. This observation is also supported by our preliminary experiments.

Table 1: The quality metrics of the considered tests, measured on the dataset of interleaving experiments (binary scheme). The values of the metrics in bold outperform other in the same column; the values marked with $\triangle$ outperform the values of the metric among other sequential tests, $p < 0.05$, Wilcoxon paired test (across folds).

Test	# stops	Type I	Type II	$Acc_{B \succ A}$	$Acc_{A \succ B}$	$\mathbb{E}(T)$, days	$\mathbb{E}(T \mid B \succ A)$	$\mathbb{E}(T \mid A \succ B)$	$\mathbb{E}(\frac{N}{N_0})$
Binomial	1	**0.00**	0.10	**0.75**	0.90	7.00	7.00	7.00	1.00
OBF-I*	7	0.01	0.10	0.73	0.92	3.17	3.17	3.04	0.44
OBF-I	7	0.01	**0.09**$^\triangle$	0.73	**0.95**$^\triangle$	3.00	3.04	2.92	0.42
MaxSPRT-I-MC	7	**0.00**$^\triangle$	0.23	0.64	0.76	3.96	4.00	3.92	0.53
MaxSPRT-I-AA	7	**0.00**$^\triangle$	0.13	0.71	0.87	3.10	3.20	3.30	0.44
OBF-I*	$7 \cdot 24$	0.01	0.11	**0.75**$^\triangle$	0.88	3.58	3.54	3.67	0.45
OBF-I	$7 \cdot 24$	0.01	**0.09**$^\triangle$	**0.75**$^\triangle$	0.93	3.33	3.38	3.29	0.44
MaxSPRT-I-MC	$7 \cdot 24$	**0.00**$^\triangle$	0.19	0.71	0.76	3.38	3.38	3.42	0.43
MaxSPRT-I-AA	$7 \cdot 24$	**0.00**$^\triangle$	0.12	0.73	0.89	**2.61**$^\triangle$	**2.63**$^\triangle$	**2.58**$^\triangle$	**0.35**$^\triangle$

with the deduped binary scheme (Table 2) the Type II error probabilities are also close for all the considered tests. However, they are considerably smaller, in the $0.03 - 0.05$ range. In both Table 1 and Table 2 the lowest Type II error is achieved by the OBF-I test. Similarly, in both cases OBF-I demonstrates the highest $Acc_{A \succ B}$ and $Acc_{B \succ A}$ metrics. In particular, in Table 1, the $Acc_{B \succ A}$ metric of 0.75 is achieved when $7 \cdot 24$ stops are used, and the $Acc_{A \succ B}$ of 0.95 is achieved when 7 stops are used.

However, when considering the mean time metric $\mathbb{E}(T)$, the difference between the tests becomes marked. In the case of the binary click aggregation scheme (Table 1), all the evaluated rules achieved considerable improvements over the standard 7-day scenario. Among the tests that use 7 stops, on average, MaxSPRT-I-MC stops the experiments later than other tests (e.g. 3.96 MaxSPRT-I-MC vs 3.10 MaxSPRT-I-AA, 7 stops). The shortest mean time (3.00) is demonstrated by the OBF-I test. Somewhat higher, but a close value of 3.10 is achieved by MaxSPRT-I-AA.

On comparing the scenarios with 7 and with $7 \cdot 24$ stops, we firstly notice that the MaxSPRT-I-MC and MaxSPRT-I-AA tests greatly benefit from using additional stops. Indeed, the mean time is reduced for the MaxSPRT-I-MC test from 3.96 to 3.38. Similarly, MaxSPRT-I-AA has improved the mean experiment running time from 3.10 to 2.61, and achieved the best performance. This behaviour is intuitive: with more stops available, there is more potential to stop earlier.

In contrast, OBF-I and OBF-I* the tests demonstrate some degradation in their $\mathbb{E}(T)$ metrics when $7 \cdot 24$ stops are used. For instance, OBF-I increased the mean deployment time from 3.00 to 3.33. A possible explanation is that OBF-I and OBF-I* rely on the central limit theorem, which only holds when the sample size approaches infinity. On the other hand, as the number of stops used by a test increases, less sessions are observed between stops, and this might harm the performance of OBF-I and OBF-I*. Another possible source of the error is that the OBF-based tests assume that the number of sessions between stops is uniform, which can be violated when stops are close.

From Table 2, we observe that MaxSPRT-AA has the shortest mean deployment time both among the tests with 7 stops (1.81), and among the tests with $7 \cdot 24$ stops (1.28). When 7 stops are used, OBF-I has a relatively close performance (1.83), but underperforms in the case of $7 \cdot 24$ stops.

Again, we notice that the OBF-I and OBF-I* tests degrade when the number of stops is increased, but MaxSPRT-I-MC and MaxSPRT-I-AA both improve their performance. Moreover, MaxSPRT-I-AA achieves the shortest mean deployment time when $7 \cdot 24$ stops are used.

On comparing Tables 1&2, we observe that when the deduped click aggregation scheme is applied, the Type II error probability and the mean deployment time decrease, while the $Acc_{A \succ B}$ and $Acc_{B \succ A}$ metrics grow for all the evaluated tests. This indicates that the deduped binary click scheme leads to marked gains in the sensitivity of the interleaving experiments.

An interesting observation is that the difference in the mean deployment times for OBF-I and OBF-I* are relatively close (not more than 0.25 days or 6 hours in the scenario with 7 stops). However, the difference between MaxSPRT-I-AA and MaxSPRT-I-MC is bigger (0.86 days ≈ 21 hours maximum). In all scenarios MaxSPRT-I-AA outperforms MaxSPRT-I-MC, indicating that replacing the Monte-Carlo threshold estimate with the threshold learned from the A/A tests improves the test's performance.

We also observe that the relative improvements measures by the $\mathbb{E}(T)$ metric are well aligned with the improvements measured by the $\mathbb{E}(\frac{N}{N_0})$ metric (e.g. MaxSPRT-I-AA with $7 \cdot 24$ stops reduces the mean deployments time by 82%, and uses only 0.15 of the available sessions).

We conclude that the MaxSPRT-I-AA test with $7 \cdot 24$ stops and the deduped binary click aggregation scheme achieves the smallest deployment time. In comparison to the standard 7-day scenario with the binary click aggregation scheme, on our dataset of the interleaving experiments, the combination of the sequential testing approach and the improved click aggregation scheme achieves 82% increase in the efficiency (1.28 vs 7.00 days).

7.2 A/B experiments

To discuss our evaluation of the sequential tests that can be applied for the A/B experiments, we use the following notation. By OBF we denote the original O'Brien&Fleming test [14], discussed in Section 4.2. Our proposed adaptation of the MaxSPRT test to the A/B experiments is further referred to as MaxSPRT-AB.

On analysing Table 3, we notice that the Type I error probabilities are close to 0.05. Again, we believe that the deviations can be explained by a limited size of the dataset. Further, we observe that the Type II error levels vary considerably among the tests: from 0.15 for the OBF test with 7 stops to 0.03 for the MaxSPRT-AB test with $7 \cdot 24$ stops.

As earlier, we observe that in terms of the mean deployment time metric MaxSPRT-AB test considerably benefits from increasing of the number of stops: $\mathbb{E}(T)$ decreases from 3.28 to 2.38. However, OBF demonstrates virtually the same mean deployment time for these two cases (4.38 vs 4.25).

Table 2: The quality metrics of the considered tests, measured on the dataset of interleaving experiments (binary deduped scheme). The values of the metrics in bold outperform other in the same column; the values marked with $\triangle$ outperform the values of the metric among other sequential tests, $p < 0.05$, Wilcoxon paired test (across folds).

Test	# stops	Type I	Type II	$Acc_{B\succ A}$	$Acc_{A\succ B}$	$\mathbb{E}(T)$, days	$\mathbb{E}(T \mid B\succ A)$	$\mathbb{E}(T \mid A\succ B)$	$\mathbb{E}(\frac{N}{N_0})$
Binomial	1	**0.00**	**0.03**	**0.82**	**0.95**	7.00	7.00	7.00	1.00
OBF-I*	7	0.01	**0.03**$^{\triangle}$	**0.82**$^{\triangle}$	**0.95**$^{\triangle}$	1.96	1.83	2.04	0.24
OBF-I	7	0.01	**0.03**$^{\triangle}$	**0.82**$^{\triangle}$	**0.95**$^{\triangle}$	1.83	1.71	2.00	0.23
MaxSPRT-I-MC	7	**0.00**$^{\triangle}$	0.05	**0.82**$^{\triangle}$	0.93	2.19	2.00	2.37	0.25
MaxSPRT-I-AA	7	**0.00**$^{\triangle}$	0.05	**0.82**$^{\triangle}$	0.93	1.81	1.64	1.98	0.22
OBF-I*	$7 \cdot 24$	0.01	0.04	**0.82**$^{\triangle}$	**0.95**$^{\triangle}$	2.13	2.04	2.21	0.26
OBF-I	$7 \cdot 24$	0.01	**0.03**$^{\triangle}$	**0.82**$^{\triangle}$	**0.95**$^{\triangle}$	1.96	1.92	2.05	0.24
MaxSPRT-I-MC	$7 \cdot 24$	**0.00**$^{\triangle}$	0.05	**0.82**$^{\triangle}$	0.93	1.63	1.44	1.85	0.17
MaxSPRT-I-AA	$7 \cdot 24$	**0.00**$^{\triangle}$	0.05	**0.82**$^{\triangle}$	0.93	**1.28**$^{\triangle}$	**1.11**$^{\triangle}$	**1.44**$^{\triangle}$	**0.15**$^{\triangle}$

Further, we notice that for any number of stops, MaxSPRT-AB demonstrates shorter mean deployment time than OBF (e.g. 4.25 vs 2.38 in the case of $7 \cdot 24$ stops). The overall minimum of $\mathbb{E}(T)$ is achieved by MaxSPRT when $7 \cdot 24$ stops are used. This value corresponds to the 66% reduction in time in comparison to the standard one-stop scenario.

7.3 Visualisation

We illustrate the best-performing MaxSPRT-I-AA and MaxSPRT-AB tests as follows. First, we sample a random subset of experiments, including experiments with A outperforming B, B outperforming A (according to the ground-truth labels), and A/A experiments. Second, for each of these experiments, at each stop i we calculate the log-likelihood ratio L_i multiplied by the sign of the current estimate of difference between A and B. More specifically, in the case of interleaving experiments, we multiply the log-likelihood (Equation (7)) by the sign of the current estimate of $(\hat{p}_1^i - \frac{1}{2})$:

$$sign\left(\hat{p}_1^i - \frac{1}{2}\right) \cdot L_i \tag{14}$$

In the case of A/B experiments, we multiply the log-likelihood ratio L_i (Equation (10)) by the sign of the differences of the current estimates of $\hat{p}_i^B - \hat{p}_i^A$:

$$sign\left(\hat{p}_i^B - \hat{p}_i^A\right) \cdot L_i \tag{15}$$

By definition, the absolute values of Equations (14)&(15) are equal to the log-likelihood ratio L_i, and their signs indicate which system tends to be better according to the observed data[4] (e.g. if (14) is positive, then $B \succ A$, and vice-versa). As a result, whenever the values of Equations (14)&(15) leave the corresponding interval $[-\bar{L}, +\bar{L}]$ ($\bar{L}$ is different for interleaving and A/B experiments), the experiment is terminated and a decision is made (e.g. $B \succ A$ if the upper boundary is touched).

We report our obtained results in Figure 1. Figures 1a and 1b correspond to the MaxSPRT-I-AA test with the binary and the binary deduped click aggregation schemes. Figure 1c corresponds to the MaxSPRT-AB test. The green and red lines correspond to the experiments that are labelled as $B \succ A$ and $A \succ B$ according to the ground-truth labels. The black lines correspond to the sampled A/A experiments. The horizontal dashed lines indicate the boundaries of the intervals $[-\bar{L}, +\bar{L}]$.

From Figure 1 we observe that despite some fluctuations, the likelihood ratios for the sampled A/A experiments are well within the boundaries of the decision interval at each of the stops. Next, for most of the interleaving experiments with A outperforming B, the statistic (14) falls towards the lower bound $(-\bar{L})$. In contrast, several experiments with $B \succ A$ have their statistic reaching the wrong boundary. These cases result in decreasing the $Acc_{B\succ A}$ value. This observation agrees with the results reported in Tables 1 & 2. Indeed, the $Acc_{B\succ A}$ metric values are lower than $Acc_{A\succ B}$ in every row of these tables. Thus, in a random sample it is more likely to meet errors of wrongly rejecting $B \succ A$ than the opposite case of rejecting $A \succ B$. Finally, on comparing Figure 1a and Figure 1b, we observe that the experiments are terminated faster in Figure 1b that corresponds to the case of the binary deduped click aggregation scheme. Again, this observation agrees with the results discussed above.

Overall, from our experiments we observe that by using sequential testing procedures considerable gains can be obtained in the mean time experiments are deployed, without significant degradation in other metrics, such as Type I, Type II errors, $Acc_{A\succ B}$, and $Acc_{B\succ A}$. In the case of the interleaving experiments with the binary click aggregation, a reduction in the execution time of 63% can be achieved by using the MaxSPRT-I-AA test with $7 \cdot 24$ stops. Moreover, the proposed sequential tests can be combined with the previously proposed deduped click aggregation scheme, and this combination results in even bigger improvements (82%). A similar observation holds in the case of the A/B tests: a decrease in the mean execution time of 66% is achieved by the MaxSPRT-AB test.

8. CONCLUSIONS AND FUTURE WORK

In this work we address an important problem of increasing the efficiency of the online evaluation experiments. In particular, we studied how sequential testing procedures can be adapted to reduce the time online evaluation experiments require. These procedures are designed so that they can stop online experiments when the observed data is sufficient to make a reliable conclusion about the experiment's outcome.

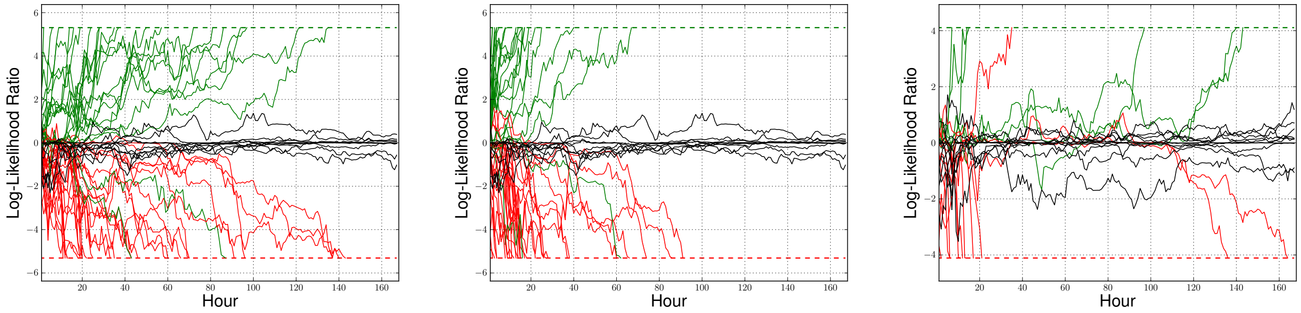
We proposed a modification of the O'Brien&Fleming group sequential test that can be applied to interleaving evaluation. Further, we described an approach to improve the MaxSPRT test's performance by adjusting its stopping threshold on the dataset of A/A experiments. Finally, we described a MaxSPRT-based test that can be applied in the A/B experiments to assess differences in the means of the abandonment rate metric.

In our evaluation study we used two datasets, containing 115 interleaving and 41 A/B experiments. Our study

[4] In the case of MaxSPRT-AB this holds as the log-likelihood ratio (10) is non-negative.

Table 3: The quality metrics of the considered tests, measured on the dataset of A/B experiments. The values of the metrics in bold outperform other in the same column; the values marked with △ outperform the values of the metric among other sequential tests in the column, $p < 0.05$, Wilcoxon paired test (across folds).

Test	# stops	Type I	Type II	$Acc_{B \succ A}$	$Acc_{A \succ B}$	$\mathbb{E}(T)$, days	$\mathbb{E}(T \mid B \succ A)$	$\mathbb{E}(T \mid A \succ B)$	$\mathbb{E}(\frac{N}{N_0})$
T-test	1	0.03	**0.03**	**1.00**	**0.93**	7.00	7.00	7.00	1.00
OBF	7	0.03	0.15	0.93△	0.70	4.38	3.90	4.65	0.60
MaxSPRT-AB	7	**0.00**△	0.11	0.87	0.78	3.28	3.48	3.17	0.43
OBF	$7 \cdot 24$	0.03	0.05	**1.00**△	0.89△	4.25	3.80	4.80	0.57
MaxSPRT-AB	$7 \cdot 24$	0.03	**0.03**△	0.93△	0.85△	**2.38**△	**2.54**△	**2.33**△	**0.31**△



(a) MaxSPRT-I-AA, binary scheme. (b) MaxSPRT-I-AA, binary deduped scheme. (c) MaxSPRT-AB, abandonment rate.

Figure 1: Illustrating the MaxSPRT tests. Green and red lines correspond to the experiments with $B \succ A$ and $A \succ B$ ground-truth labels, respectively. Black lines correspond to A/A experiments. The horizontal dashed lines denote the threshold values for accepting $B \succ A$ (green) and $A \succ B$ (red).

demonstrates that by using the sequential testing procedures, a marked reduction in the duration of the experiments can be achieved, without significant losses in other metrics, such as Type I and Type II error probabilities. The maximal improvement over a standard 7 day one-stop scenario on the dataset of interleaving experiments reaches 63% by using the MaxSPRT-I-AA test, which examines the experiment data every hour. Further improvement can be obtained by additionally using an improved deduped binary click aggregation scheme, and it reaches 82%. This supports the idea that the sequential testing approach is complimentary to the previous research, which concentrated on reducing noise in the user feedback. On the dataset of A/B experiments, the MaxSPRT-AB test obtains an improvement of 66% over the same standard evaluation scenario. An interesting direction of future work is to devise a MaxSPRT-based test for the non-binomial A/B metrics (e.g., absence time [5]).

9. REFERENCES

[1] K. Balog, L. Kelly, and A. Schuth. Head first: Living labs for ad-hoc search evaluation. In *CIKM 2014*.

[2] J. Bartroff, T. L. Lai, and M.-C. Shih. *Sequential Experimentation in Clinical Trials: Design and Analysis*, volume 298. Springer, 2012.

[3] O. Chapelle, T. Joachims, F. Radlinski, and Y. Yue. Large-scale validation and analysis of interleaved search evaluation. *ACM TOIS*, 30(1):6, 2012.

[4] A. Deng, Y. Xu, R. Kohavi, and T. Walker. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *WSDM 2013*.

[5] G. Dupret and M. Lalmas. Absence time and user engagement: evaluating ranking functions. In *WSDM 2013*.

[6] K. Hofmann, S. Whiteson, and M. de Rijke. A probabilistic method for inferring preferences from clicks. In *CIKM 2011*.

[7] T. Joachims. Optimizing search engines using clickthrough data. In *KDD 2002*.

[8] E. Kharitonov, C. Macdonald, P. Serdyukov, and I. Ounis. Using historical click data to increase interleaving sensitivity. In *CIKM 2013*.

[9] R. Kohavi, T. Crook, R. Longbotham, B. Frasca, R. Henne, J. L. Ferres, and T. Melamed. Online experimentation at microsoft. *Data Mining Case Studies*, page 11, 2009.

[10] R. Kohavi, A. Deng, B. Frasca, R. Longbotham, T. Walker, and Y. Xu. Trustworthy online controlled experiments: Five puzzling outcomes explained. In *KDD 2012*.

[11] R. Kohavi, A. Deng, B. Frasca, T. Walker, Y. Xu, and N. Pohlmann. Online controlled experiments at large scale. In *KDD 2013*.

[12] M. Kulldorff, R. L. Davis, M. Kolczak, E. Lewis, T. Lieu, and R. Platt. A maximized sequential probability ratio test for drug and vaccine safety surveillance. *Sequential Analysis*, 30(1):58–78, 2011.

[13] K. G. Lan and D. L. DeMets. Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663, 1983.

[14] P. C. O'Brien and T. R. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, 1979.

[15] S. J. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 1977.

[16] F. Radlinski and N. Craswell. Comparing the sensitivity of information retrieval metrics. In *SIGIR 2010*.

[17] F. Radlinski and N. Craswell. Optimized interleaving for online retrieval evaluation. In *WSDM 2013*.

[18] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *CIKM 2008*.

[19] D. Siegmund. *Sequential analysis: tests and confidence intervals*. Springer, 1985.

[20] A. Wald. Sequential tests of statistical hypotheses. *Ann. Math. Statist.*, 16(2):117–186, 06 1945.

[21] A. Wald and J. Wolfowitz. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, pages 326–339, 1948.

[22] K. Wang, T. Walker, and Z. Zheng. Pskip: estimating relevance ranking quality from web search clickthrough data. In *SIGKDD 2009*.

[23] S. K. Wang and A. A. Tsiatis. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, 1987.

[24] Y. Yue, Y. Gao, O. Chapelle, Y. Zhang, and T. Joachims. Learning more powerful test statistics for click-based retrieval evaluation. In *SIGIR 2010*.