A Nonparametric Sequential Test for Online Randomized Experiments

Vineet Abhishek
Walmart Labs
Sunnyvale, CA, USA
vabhishek@walmartlabs.com

Shie Mannor
Technion
Haifa, Israel
shie@ee.technion.ac.il

ABSTRACT

We propose a nonparametric sequential test that aims to address two practical problems pertinent to online randomized experiments: (i) how to do a hypothesis test for complex metrics; (ii) how to prevent type 1 error inflation under continuous monitoring. The proposed test does not require knowledge of the underlying probability distribution generating the data. We use the bootstrap to estimate the likelihood for blocks of data followed by mixture sequential probability ratio test. We validate this procedure on data from a major online e-commerce website. We show that the proposed test controls type 1 error at any time, has good power, is robust to misspecification in the distribution generating the data, and allows quick inference in online randomized experiments.

Keywords

Sequential tests; the bootstrap; A/B tests

1. INTRODUCTION

Most major online websites continuously run several randomized experiments to measure the performance of new features or algorithms and to determine their successes or failures; see [10] for a comprehensive survey. A typical workflow consists of: (i) defining a success metric; (ii) performing a statistical hypothesis test to determine whether or not the observed change in the success metric is significant. Both of these steps, however, are nontrivial in practice.

Consider a typical product search performance data collected by any online e-commerce website. Such data is multidimensional, containing dimensions such as the number of search queries, the number of conversions, time to the first click, total revenue from the sale, etc., for each user session. The success metric of interest could be a complicated scalar function of this multidimensional data rather than some simple average; e.g., a metric of interest could be the ratio of expected number of successful queries per user session to the expected number of total queries per user session, an empirical estimate of which is a ratio of two sums. Furthermore, some commonly used families of probability distributions such as the Bernoulli distribution or the normal distribution are almost never suitable to model this data. The search queries in a user session are highly correlated; using the Bernoulli distribution to model even simple metrics such the click through rate per impression is inaccurate.[1] Another example is the revenue based metrics with a considerable mass at zero and a heavy tail distribution for the nonzero values. The use of the normal distribution for modeling revenue per user session is inaccurate and sensitive to extreme values unless the sample size is huge. As a result, the use of the binomial proportion test or the t-test for hypothesis testing is unsuitable for such metrics respectively. It requires considerable effort to find appropriate models for each of these metrics and it is implausible to come up with a suitable parametrized family of probability distributions that could be used to model different complex success metrics.

Additionally, classical statistical tests such as the t-test assume that the number of samples is given apriori. This assumption easily breaks in an online setting where an end user with not much formal training in statistics has near real-time access to the performance of ongoing randomized experiments: the decision to stop or continue collecting more data is often based on whether or not the statistical test on the data collected so far shows significance or not. This practice of chasing significance leads to highly inflated type 1 error.[2] Several recent articles have highlighted this aspect of data collection as one of the factors contributing to the statistical crisis in science; see, e.g., [18] and [5]. Use of sample size calculator to determine the number of samples to be collected and waiting until the required number of samples have been collected is hard to enforce in practice, especially in do-it-yourself type platforms commonly used to set up and monitor online randomized experiments. Additionally, the sample size calculation requires an estimate of the minimum effect size and the variance of the metric. These estimates are often crude and result in waiting longer than necessary in case the true change is much larger than the estimate.

The above two challenges point to two separate fields of statistics: the bootstrap and sequential tests. The bootstrap is frequently used to estimate the properties of complicated

©2017 International World Wide Web Conference Committee (IW3C2), published under Creative Commons CC BY 4.0 License.
WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054196

[1] A similar observation is made by [1]. The authors suggest avoiding the use of standard formulas for standard deviation computation which assume independence of the samples and recommend the use of the bootstrap.

[2] The inflation of type 1 error because of data collection to reach significance can be explained by the law of the iterated logarithm, as observed in [14]. Section 5 provides some plots to demonstrate the ill effects of chasing significance.

statistics; see [4] and [7] for a comprehensive review. Sequential tests provide an alternative to committing to a fixed sample size by providing type 1 error control at any time under continuous monitoring; see [12] for a survey on sequential tests. Sequential tests, however, are still uncommon in practice partly because of the information they require about the data generation process.

Our main contribution is in bringing together the ideas from the bootstrap and sequential tests and applying them to construct a nonparametric sequential test for performing a hypothesis test on complex metrics. We provide empirical evidence using the data from a major online e-commerce website demonstrating that our nonparametric sequential test is suitable for the diverse type of metrics, ensures that type 1 error is controlled at any time under continuous monitoring, has good power, is robust to misspecification in the distribution generating the data, and enables making quick inference in online randomized experiments. We advocate the use of this general approach in the absence of knowledge of the right distribution generating the data, which typically is the case.

Briefly, we compute a studentized statistic for the metric of interest on blocks of data and use the bootstrap to estimate its likelihood in a nonparametric manner. The likelihood ratios can then be used to perform mixture sequential probability ratio test (henceforth, mixture SPRT) [17]. We highlight theoretical results from existing literature justifying the use of the bootstrap based likelihood estimation and the mixture SPRT.

Related work: The bootstrap based nonparametric approach we take distinguishes our work from the related literature on sequential testing. Most of the existing literature assumes the knowledge of parametrized family of distribution from which data (typically scalar) is generated; ours does not. The idea of sequential probability ratio test (SPRT) for testing a simple hypothesis and its optimality originates from the seminal work of [20]. Mixture SPRT [17], described in Section 3.2, extends this to a composite hypothesis. For the exponential family of distributions, mixture SPRT is shown to be a test with power 1 and almost optimal with respect to expected time to stop; see [17] and [16]. [8] formalizes the notion of always valid p-value and uses mixture SPRT for the sequential test. MaxSPRT [11], [9] is another popular sequential test where the unknown parameter is replaced by its maximum likelihood estimate computed from the observed data so far and the rejection threshold is determined (typically using numerical simulations) to ensure type 1 error control. The computation of the maximum likelihood estimate as well as the rejection threshold requires knowing the parametrized family of distribution; consequently, MaxSPRT has been studied only for a few simple distributions. Group sequential tests [15] divide observations into blocks of data for inference. It, however, assumes the normal distribution and uses mean as the metric of interest. The theoretical foundations of the bootstrap likelihood estimation can be found in [6]; it shows that the bootstrap based likelihood estimation of the studentized statistic provide a fast rate of convergence to the true likelihood. [2] uses a nested bootstrap to compute the likelihood which can be computationally expensive. [3] uses confidence intervals to compute the likelihood and shows that for the exponential family of distributions, the confidence interval based approach has the rate of convergence of the same order as in [6].

2. MODEL AND ASSUMPTIONS

Let $\mathbf{X} \in \mathbb{R}^d$ be a random vector with the CDF $F(\mathbf{x}; \theta, \eta)$ and corresponding pdf $f(\mathbf{x}; \theta, \eta)$. Here θ is the scalar parameter of interest and η is a nuisance parameter (possibly null or multidimensional). Denote a realized value of the random vector $\mathbf{X}$ by $\mathbf{x}$. For $j < k$, let $\mathbf{X}_{j:k} \triangleq (\mathbf{X}_j, \mathbf{X}_{j+1}, \ldots, \mathbf{X}_k)$ be a collection of i.i.d. random vectors indexed from j to k; define $\mathbf{x}_{j:k}$ similarly.

We are interested in testing the null hypothesis that $\theta = \theta_0$ against the alternate hypothesis $\theta \neq \theta_0$ such that type 1 error is controlled at any given time and the power of the test should be high; i.e., we are looking for a function D_{θ_0} mapping sample paths $\mathbf{x}_1, \mathbf{x}_2, \ldots$ to $[0, 1]$ such that for any $\alpha \in (0, 1)$:

$$\mathbb{P}_\theta \left[D_{\theta_0}(\mathbf{X}_{1:n}) \leq \alpha \text{ for some } n \geq 1 \right] \quad \begin{cases} \leq \alpha & \text{if } \theta = \theta_0, \\ \approx 1 & \text{if } \theta \neq \theta_0, \end{cases} \tag{1}$$

where $\mathbf{X}$'s are assumed to be realized i.i.d. from $F(\mathbf{x}; \theta, \eta)$ and $\mathbb{P}_\theta$ is the probability under $F(\mathbf{x}; \theta, \eta)$. The first part of (1) is a statement on type 1 error control: if the null hypothesis is true and our tolerance for incorrect rejection of the null hypothesis is α, then the probability that the test statistic exceeds α and rejects the null hypothesis (*type 1 error*) is at most α. The second part of (1) is a statement on the power of the test: if the null hypothesis is false then the probability that the test statistic exceeds α and rejects the null hypothesis (*the power of the test*) is close to one.

Let $\widehat{F}(\mathbf{x}_{1:n})$ be the empirical CDF constructed from n i.i.d. samples. We make the following assumption about the parameter of interest θ:

Assumption 1. Assume that θ is a functional statistic with asymptotically normal distribution, i.e., $\theta = T(F)$ for some functional T and $\sqrt{n}(T(\widehat{F}(\mathbf{X}_{1:n})) - \theta) \to \mathcal{N}(0, \zeta^2)$ in distribution as $n \to \infty$, where $T(\widehat{F}(\mathbf{x}_{1:n}))$ is the plug-in estimate of θ. Furthermore, assume that the asymptotic variance ζ^2 does not depend on the specific choice of θ.

Asymptotic normality of plug-in statistic in Assumption 1 holds under fairly general conditions and captures several statistics of interest; see [13] for a formal discussion on this.

3. PRELIMINARIES

Our work leverages ideas from the bootstrap and mixture SPRT. This section provides their brief overview.

3.1 The bootstrap

Let $Z_{1:n}$ be n i.i.d. random variables and let $H_n(z) \triangleq \mathbb{P}\left[T(Z_{1:n}) \leq z\right]$ be the sampling distribution of a functional T. The CDF of Z's is usually unknown and hence H_n cannot be computed directly. The (nonparametric) bootstrap is used to approximate the sampling distribution H_n using a resampling technique. Let $Z^*_{1:n}$ be n i.i.d. random variables each with the CDF equal to the empirical CDF constructed from $Z_{1:n}$. Let $H^*_n(z) \triangleq \mathbb{P}\left[T(Z^*_{1:n}) \leq z \mid Z_{1:n}\right]$ be the bootstrap distribution of the functional T. Notice that H^*_n is a random CDF that depends on the realized values of $Z_{1:n}$. The idea is to use H^*_n as an approximation for H_n.[3]

[3]Under some regularity conditions and a suitable distance metric on the space of CDFs, distance between H_n and H^*_n goes to zero with probability 1 as $n \to \infty$; see [7].

The bootstrap distribution is computed using Monte Carlo method by drawing a large number of samples, each of size n, from the empirical distribution. This is identical to sampling with replacement and provides a convenient computational method to deal with complex statistics. Even in the cases where the statistics are asymptotically normal, we might have to rely on the bootstrap to compute its asymptotic variance. Moreover, for asymptotically normal statistics, the bootstrap typically provides an order improvement in convergence rate to the true distribution (see [7] and [6]). This motivates the use of the bootstrap in our nonparametric sequential test.

3.2 Mixture SPRT

In a general form, mixture SRPT works as follows. Let $\rho_n(z_{1:n};\theta)$ be the joint pdf parametrized by θ of a sequence of random variables $Z_{1:n}$. Let $\theta \in \mathcal{I}$ and $\pi(\theta)$ be any prior on θ with $\pi(\theta) > 0$ for $\theta \in \mathcal{I}$. Define:

$$L(z_{1:n};\theta_0) \triangleq \frac{\int_{\theta\in\mathcal{I}} \rho_n(z_{1:n};\theta)\pi(\theta)}{\rho_n(z_{1:n};\theta_0)}. \tag{2}$$

If $\theta = \theta_0$, the sequence $L(Z_{1:n};\theta_0)$'s forms a martingale with respect to the filtration induced by the sequence Z_n's.[4] Using Doob's martingale inequality,

$$\mathbb{P}_{\theta_0}\left[L(Z_{1:n};\theta_0) > \frac{1}{\alpha} \text{ for some } n \geq 1\right] \leq \alpha \text{ for any } \alpha > 0, \tag{3}$$

where the probability is computed assuming $\theta = \theta_0$. An immediate consequence of this is control of type 1 error at any time. The optimality of mixture SPRT with respect of power and the expected time to stop ([17] and [16]) and its ease of implementation motivates its use in our nonparametric sequential test.

The p-value at time n is defined as:

$$p_{value}(z_{1:n};\theta_0) \triangleq \min\left\{1, \frac{1}{\max\{L(z_{1:t};\theta_0) : t \leq n\}}\right\}. \tag{4}$$

It follows immediate from (3) and (4) that

$$\mathbb{P}_{\theta_0}\left[p_{value}(Z_{1:n};\theta_0) \leq \alpha\right] \leq \alpha \quad \text{for any } n \tag{5}$$

and $p_{value}(z_{1:n};\theta_0)$ is nonincreasing in n. In this sense, this is an always valid p-value sequence, defined in [8]. The duality between the p-value and the confidence interval can be used to construct an always valid (though possibly larger than the exact) confidence interval.

Mixture SPRT, however, requires the knowledge of the functional form of the pdf up to the unknown parameter of interest. This is hard especially in a multidimensional setting and in the presence of nuisance parameter(s). We address this using the bootstrap, as described next.

4. THE NONPARAMETRIC SEQUENTIAL TEST

This section shows how mixture SPRT and the bootstrap could be used together to construct a nonparametric sequential test for complex statistics. We will also discuss how to apply the ideas developed here to A/B tests.

[4]This follows from a general observation that if $Z_{1:n} \sim \rho_n(z_{1:n})$ and $\psi_n(z_{1:n})$ is any other probability density function, then $L_n = {}^{\psi_n(z_{1:n})}/_{\rho_n(z_{1:n})}$ is a martingale.

4.1 The bootstrap mixture SPRT

Our main idea is as follows: (i) divide the data into blocks of size N; (ii) compute a studentized plug-in statistic for θ on each block of data; (iii) estimate the likelihood of the studentized plug-in statistic for each block using the bootstrap; (iv) use mixture SPRT by treating each block as a unit observation. We provide details below.

Let $\mathbf{x}_{(k)} \triangleq \mathbf{x}_{kN:(k+1)N-1}$ denote the kth block of data of size N and $s(\mathbf{x}_{(k)};\theta)$ be the studentized plug-in statistic computed from the kth data block, i.e.,

$$s(\mathbf{x}^{(k)};\theta) = \frac{T(\widehat{F}(\mathbf{x}_{(k)})) - \theta}{\sigma(T(\widehat{F}(\mathbf{x}_{(k)})))}, \tag{6}$$

where $\sigma(T(\widehat{F}(\mathbf{x}_{(k)})))$ is a consistent estimate of the standard deviation of the plug-in statistic $T(\widehat{F}(\mathbf{X}_{(k)}))$.[5] For notational convenience, we use $\widehat{\theta}_{(k)}$ and $s^{\theta}_{(k)}$ in lieu of $T(\widehat{F}(\mathbf{x}_{(k)}))$ and $s(\mathbf{x}_{(k)};\theta)$ respectively, with the dependence on underlying $\mathbf{x}$'s and N being implicit; and their uppercase versions to denote the corresponding random variables. The random variable $S^{\theta}_{(k)}$ is asymptotically distributed as $\mathcal{N}(0,1)$ at rate $O({}^{1}/\sqrt{N})$ assuming θ is the true parameter. Hence for large N, the distribution of $S^{\theta}_{(k)}$ is approximately pivotal (i.e., the distribution does not depend on the unknown θ and the effect of η is negligible). This suggests using the normal density as the pdf for $S^{\theta}_{(k)}$ and carry out mixture SPRT following the steps in Section 3.2. The density, however, can be approximated better using the bootstrap.

Algorithm 1 below describes the bootstrap mixture SPRT.

Correctness of Algorithm 1: We give a heuristic explanation of the correctness of Algorithm 1; a rigorous proof of type 1 error control or optimality of Algorithm 1 in some sense is extremely hard.[6] L_n in step 3 is similar to (2). If L_n is approximately a martingale for $\theta = \theta_0$, the arguments of Section 3.2 would imply type 1 error control at any time under continuous monitoring. Step 2 of Algorithm 1 is approximating the true pdf of the random variable $S^{\theta}_{(k)}$, denoted by $g^{\theta}_{(k)}$, using the bootstrap based estimate $g^{*}_{(k)}$. We only need to ensure that $g^{*}_{(k)}$ is a good approximation of $g^{\theta_0}_{(k)}$, assuming $\theta = \theta_0$, for the martingale property to hold. A subtlety arises because the bootstrap distribution is discrete. This, however, is not a problem: [6] shows that a smoothened version of bootstrap density obtained by adding a small independent noise $\mathcal{N}(0, {}^{1}/N^{c})$ with arbitrary large c to the bootstrap samples converges uniformly to the true density at rate $O_p({}^{1}/N)$. Working with the studentized statistic is crucial in obtaining $O_p({}^{1}/N)$ convergence; see [7] for further details. By contrast, approximating $g_{(k)}$ by the normal density is accurate to $O({}^{1}/\sqrt{N})$. The bootstrap samples behave as if realized from a true density. A Gaussian kernel density can be used to estimate the bootstrap density from a finite number of bootstrap samples, as in step 2 of Algorithm 1. The independence of the asymptotic variance and θ by Assumption 1 allows us to compute the likelihood for

[5]Formally, $\sqrt{n}\sigma(T(\widehat{F}(\mathbf{X}_{1:n}))) \to \zeta$ with probability 1 as $n \to \infty$, where ζ is defined in Assumption 1.

[6]The results on the optimality of mixture SPRT are known only for exponential families of distributions. We, however, take a nonparametric approach. A theory of bootstrap is based on Edgeworth expansion [7]. Applying Edgeworth expansion to analyze likelihood ratio statistic defined in (8) is quite challenging.

Algorithm 1 The bootstrap mixture SPRT

Given blocks of observations $\mathbf{x}_{(k)}$ for $k = 1, 2, \ldots$, and a prior $\pi(\theta)$:

1. Draw $\widetilde{\theta}_1, \widetilde{\theta}_2, \ldots \widetilde{\theta}_M$ i.i.d. from $\pi(\theta)$ for some large M.
2. For each k:
   (a) Compute $\widehat{\theta}_{(k)}$ and its standard deviation $\sigma(\widehat{\theta}_{(k)})$.
   (b) Draw a random sample of size N with replacement from $\mathbf{x}_{(k)}$; denote it by $\mathbf{x}^*_{(k)}$. Repeat this B times to obtain a sequence $\mathbf{x}^{*b}_{(k)}$, $1 \leq b \leq B$. For each b, compute the studentized plug-in statistic $s^{*b}_{(k)}$ as:
   $$s^{*b}_{(k)} = \frac{T(\widehat{F}(\mathbf{x}^{*b}_{(k)})) - \widehat{\theta}_{(k)}}{\sigma(T(\widehat{F}(\mathbf{x}^{*b}_{(k)})))}. \quad (7)$$
   Let $g^*_{(k)}(\cdot)$ be the pdf obtained using a Gaussian kernel density estimate from the sequence $s^{*b}_{(k)}$, $1 \leq b \leq B$.
3. Compute L_n as follows:
   $$L_n = \sum_{m=1}^{M} \left(\frac{\prod_{k=1}^{n} g^*_{(k)} \left(\frac{\widehat{\theta}_{(k)} - \widetilde{\theta}_m}{\sigma(\widehat{\theta}_{(k)})} \right)}{\prod_{k=1}^{n} g^*_{(k)} \left(\frac{\widehat{\theta}_{(k)} - \theta_0}{\sigma(\widehat{\theta}_{(k)})} \right)} \right). \quad (8)$$
4. Reject the null hypothesis $\theta = \theta_0$ if $L_n > 1/\alpha$ for some $n > 1$.

different values of θ. Step 4 is a Monte Carlo computation of Bayesian average in the right side of (2).

4.2 Application to A/B tests

In A/B tests, the goal is to identify whether or not the success metric computed on each of the two groups of data has a nonzero difference (i.e., $\theta_0 = 0$). The bootstrap mixture SPRT easily extends to A/B tests. Assume for simplicity that the block size is the same for both groups. Let $\mathbf{x}_{(k)}$'s be data from group A (control) and $\mathbf{y}_{(k)}$'s the data from group B (variation). We modify the quantities $\widehat{\theta}_{(k)}$, $\sigma(\theta_{(k)})$, and $s^{*b}_{(k)}$ of Algorithm 1 as follows:

$$\widehat{\theta}_{(k)} = T(\widehat{F}(\mathbf{y}_{(k)})) - T(\widehat{F}(\mathbf{x}_{(k)})),$$
$$\sigma(\widehat{\theta}_{(k)}) = \sqrt{\sigma^2(T(\widehat{F}(\mathbf{y}_{(k)}))) + \sigma^2(T(\widehat{F}(\mathbf{x}_{(k)})))},$$
$$s^{*b}_{(k)} = \frac{T(\widehat{F}(\mathbf{y}^{*b}_{(k)})) - T(\widehat{F}(\mathbf{x}^{*b}_{(k)})) - \widehat{\theta}_{(k)}}{\sqrt{\sigma^2(T(\widehat{F}(\mathbf{x}^{*b}_{(k)}))) + \sigma^2(T(\widehat{F}(\mathbf{x}^{*b}_{(k)})))}}.$$

We conclude with a discussion on how to determine various parameters in Algorithm 1. Type 1 error control holds for any prior $\pi(\theta)$ with nonzero density on possible values of θ. The choice of prior, however, can affect the number of observations needed to reject the null hypothesis. A simple choice is $\mathcal{N}(0, \tau^2)$; τ can be determined from the outcomes of the past A/B tests or can be set to a few percent of the reference value, computed from the data collected before the start of the A/B test, of the metric of interest. The standard deviation in (7) for each b can be computed using the bootstrap again on $\mathbf{x}^{*b}_{(k)}$. This, however, is computationally prohibitive. In practice, the standard deviation is computed using the delta method or the jackknife; see e.g., [4]. A similar method can be used to compute the standard deviation in (6). The parameters B and M are merely used for Monte Carlo integration; any sufficiently large value is good. Typical choices for the number of bootstrap sample B is $1000 - 5000$ and the number of Monte Carlo samples M is $5000 - 10,000$. Because we work with a block of data at a time, the computational overhead is not much; computation can be performed in an online manner as new blocks of data arrive or using a MapReduce approach on the historical data. Finally, a small block size is intuitively expected to enable fast inference. Doob's martingale inequality is far from being tight for martingales stopped at some maximum length and becomes tighter as this maximum length is increased. However, the small block size may reduce the accuracy of the bootstrap likelihood estimate. Additionally, data update frequency of a near real-time system puts a lower limit on how small the block of data can be; it is not uncommon to receive a few hundred or a few thousand data points per update. Section 5 describes how to estimate a reasonable block size using post A/A tests.

5. EMPIRICAL EVALUATION AND DISCUSSION

Datasets, metrics, and baselines: We use data obtained from a major online e-commerce website on product search performance. Each dataset contains a tiny fraction of randomly chosen anonymized users; and for each user session, it contains the session timestamp, the number of distinct search queries, the number of successful search queries, and the total revenue. Datasets contain about 1.8 million user sessions each and span nonoverlapping 2-3 weeks timeframe. We show here the results for only one dataset; similar results were observed for the other datasets.

We work with two metrics: (i) *query success rate per user session*; (ii) *revenue per user session.* The former is the ratio of the sum of successful queries per user session to the sum of total queries per user session, where the sum is over user sessions. The latter is a simple average, however, with most values being zero and nonzero values showing a heavy tail. These two very different metrics are chosen to evaluate the flexibility of the bootstrap mixture SPRT. We have used $\mathcal{N}(0, \tau^2)$ distribution as prior $\pi(\theta)$, with τ being equal to a few percent (less than 5%) of the values of metrics computed on data before the start timestamp of the dataset under consideration. Data is partitioned into blocks based on the order induced by the user session timestamp.

Existing sequential tests require knowing a parameterized distribution generating the data. The data generation model for the two metrics we study in our experiments is hard to find. We are not aware of any existing sequential tests that can be applied to the two complex metrics we use in our experiments. Consequently, we use the z-test on the entire data as a baseline for comparison.[7] The use of z-test (or the t-test) is pervasive in commercial A/B tests platforms, often without much scrutiny. Additionally, we compare the bootstrap mixture SPRT with MaxSPRT ([11], [9]) on synthetic data generated from the Bernoulli distribution. MaxSPRT requires knowing a parametrized likelihood function and has

[7]Because of the large size of the dataset we use, there is no observable difference between the t-test and the z-test.

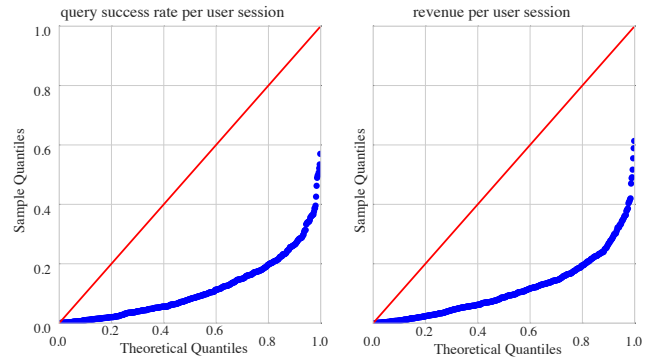
been applied mostly when the data is generated from simple distributions such as the Poisson distribution and the Bernoulli distribution.

Post A/A tests based methodology: We make use of post A/A tests to estimate the type 1 error. For each dataset, we randomly divide the users into two groups, perform the statistical test for the null hypothesis of zero difference in the metrics for the two groups, and compute the p-values. This process is repeated multiple times (250 in our experiments). Because there is no true difference between the two groups, the p-value should theoretically follow uniform distribution between 0 and 1. Consequently, the Q-Q plots for the realized p-values from different random splits and the uniform distribution should be the line of slope 1. To estimate the power of the test, for each random split of the dataset, we add to the realized difference in the metrics for the two groups some small positive offset. For each choice of the additive offset, we then check how many times the null hypothesis is rejected using p-value ≤ 0.05 as our null hypothesis rejection criteria.

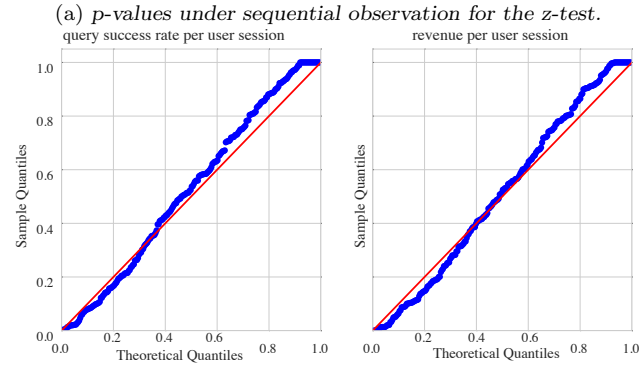
Type 1 error: Figure 1a shows a huge inflation of type 1 error under sequential monitoring in case of classical fixed sample size z-test. The Q-Q plot lies well below the line of slope 1. Here, the p-values are computed daily using the data collected until that day. The Q-Q plot is for the minimum p-value observed from daily observations, modeling the scenario where one waits until the data shows significance. Waiting until the p-value is ≤ 0.05 and stopping the test on that day resulted in type 1 error rate of at least 30%. By contrast, the Q-Q plot of always valid p-value in the case of the bootstrap mixture SPRT with the block size 5000 is close to the line of slope 1, as shown in Figure 1b. This implies that the bootstrap mixture SPRT controls type 1 error at any time under continuous monitoring. The p-value is the final one at the end of all the blocks (recall that the p-value is nonincreasing in this case).

The choice of the block size of 5000 is determined by post A/A tests with different block sizes, looking at the corresponding Q-Q plots, and making sure it stays above but close to the line of slope 1. Smaller block size makes the bootstrap approximation inaccurate while the larger block size makes this procedure more conservative in case of some upper limit on the number of samples available for the test. This is a consequence of fewer intermediate observations in a finite length dataset, resulting in Doob's martingale inequality far from being tight. Figure 2a and 2b show the Q-Q plot for block length of $10,000$ and $20,000$ respectively. The plots lie above the line of slope 1, showing a more conservative type 1 error control.

Test power and average duration: Next, we compare the power of the bootstrap mixture SPRT with the z-test on the entire data. We use the delta method to compute the standard deviation for the query success rate; see [19]. We use an additive offset to the realized difference, as described earlier in this section. Figure 3a shows the power of the test as a function of the true difference percentage for the case of query success rate. Here the true difference percentage is the additive offset normalized by the value of the metric on the entire dataset. Notice some loss in power compared to the z-test on the entire dataset. This loss in power is expected because of the sequential nature of the procedure on a finite data set; equivalently, the bootstrap mixture SPRT requires a larger number of samples to achieve the same power as the



(a) *p-values under sequential observation for the z-test.*



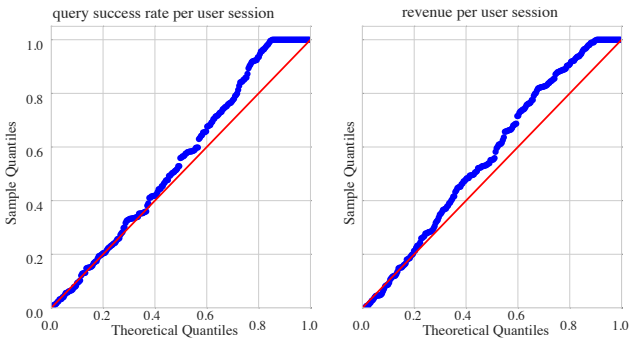
(b) *p-values under the bootstrap mixture SPRT.*

Figure 1: Q-Q plot of p-values computed on random splits of the dataset.

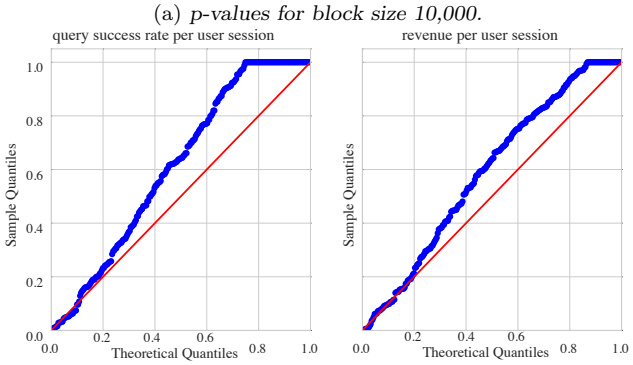
fixed sample size tests. This is the price paid for the flexibility of continuous monitoring. The loss in power quickly becomes negligible as the true difference becomes large. Focusing on the revenue metric, we notice something surprising: the z-test on the entire data has smaller power than the bootstrap mixture SPRT, as shown in Figure 3b. In this case, the mean of data does not confirm to the normal distribution used to justify the z-test. This is illustrated by the z-test on the entire data after outliers removal which has higher power than the bootstrap mixture SPRT, as expected from a test with an apriori access to the entire data.[8] The robustness of the bootstrap mixture SPRT to outliers comes from two factors: (i) the use of the bootstrap for learning the likelihood which typically works better for the heavy tail distributions; (ii) working with blocks of data confines the effect of extreme values in the data in a block to that block only by making the likelihood ratio for that block close to one without affecting the other blocks. Thus, not only the bootstrap mixture SPRT allows continuous monitoring, it is more robust to the misspecification in the distribution generating the data.

Figure 4 shows the average duration of the test using the bootstrap SPRT. The test duration is the numbers of samples consumed in each split until either all the data is exhausted or the null hypothesis is rejected. We take the average over the 250 random splits of the data set. Notice

[8]Identifying outliers is a complicated problem by itself, often with no good agreement on the definition of outliers. Here, we simply clip the values to some upper limit affecting only a negligible portion of data to show the effect of extreme values in data on the z-test.

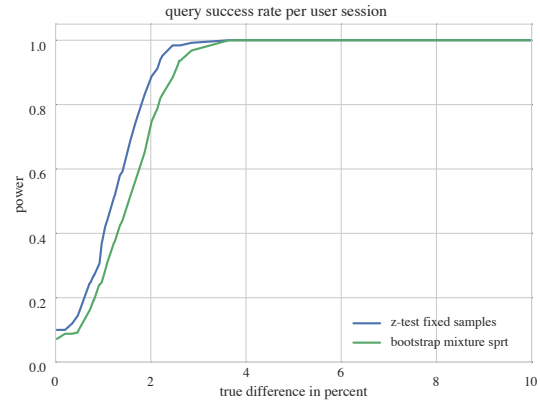


(a) *p-values for block size 10,000.*

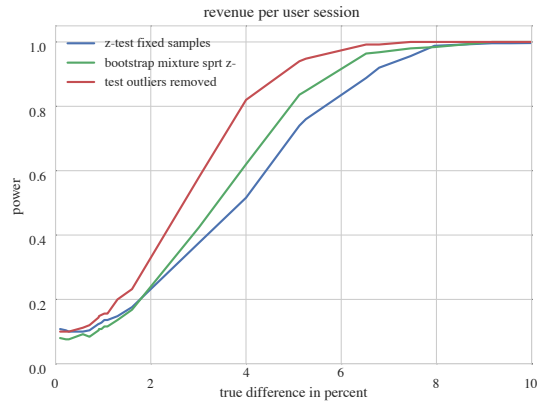


(b) *p-values for block size 20,000.*

Figure 2: Q-Q plot of p-values under bootstrap mixture SPRT with different block sizes.



(a) *test power for query success rate per user session.*



(b) *test power for revenue success rate per user session.*

Figure 3: Power comparison for z-test with fixed sample size and the bootstrap mixture SPRT.

that the null hypothesis is rejected quickly for large changes, unlike the fixed sample size test where one has to wait until the precomputed number of samples have been collected. This avoids wasteful data collection and unnecessary wait in case of misspecification in the minimum effect size used to determine the number of samples to be collected for fixed sample size tests.

In our experiments, we observed the power of the tests with smaller block size to be either comparable or better than the power of the tests with larger block size; this is in-line with the observations made earlier about the advantage of the small block size. Similarly, the tests with smaller block size had smaller test duration.

Comparison with MaxSPRT on synthetic data: Finally, we compare the performance of the bootstrap mixture SPRT with MaxSPRT. We use synthetic data of size $100,000$ generated from the Bernoulli distribution with success probability equal to 0.05. The rejection threshold for MaxSPRT is determined by multiple A/A tests on synthetic datasets such that the type 1 error is close to 0.05, as described in [9]. We also use A/A tests to determine the block size and the variance of the Gaussian prior for the bootstrap SPRT such that the type 1 is close to 0.05. This resulted in the block size of 1000. Figure 5a shows comparable test power for MaxSPRT and the bootstrap mixture SPRT with MaxSPRT performing marginally better. The better performance of MaxSPRT is expected; here the data distribution is exactly known and is exploited by MaxSPRT. The bootstrap mixture SPRT, however, learns the distribution in a nonparametric way and yet performs comparable to MaxSPRT. A similar observation can be made for the average test duration: MaxSPRT has slightly smaller test duration as shown in Figure 5b. MaxSPRT updates its likelihood ratio score for each data point as it arrives. By contrast the bootstrap mixture SPRT updates its likelihood ratio score in blocks of data which contributes to a slightly longer average test duration. Block update, however, is a more realistic scenario in a practical system where data usually arrives in blocks of size few hundreds to a few thousand. The general purpose nature of the bootstrap mixture SPRT while still performing comparable to MaxSPRT which requires exact knowledge of the data distribution is appealing.

6. CONCLUSIONS

We propose the bootstrap mixture SPRT for the hypothesis test of complex metrics in a sequential manner. We highlight the theoretical results justifying our approach and provide empirical evidence demonstrating its virtues using real data from an e-commerce website. The bootstrap mixture SPRT allows the flexibility of continuous monitoring while controlling the type 1 error at any time, has good power, is robust to misspecification in the distribution generating the data, and allows quick inference in case the true change is larger than the estimated minimum effect size. The nonparametric nature of this test makes it suitable for the cases where the

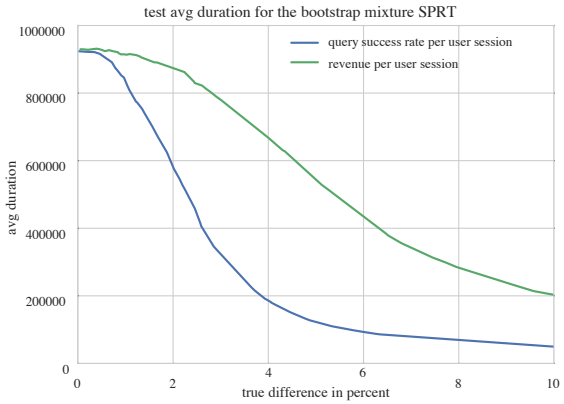
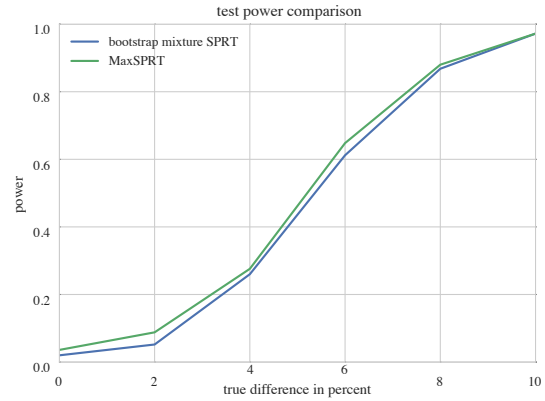
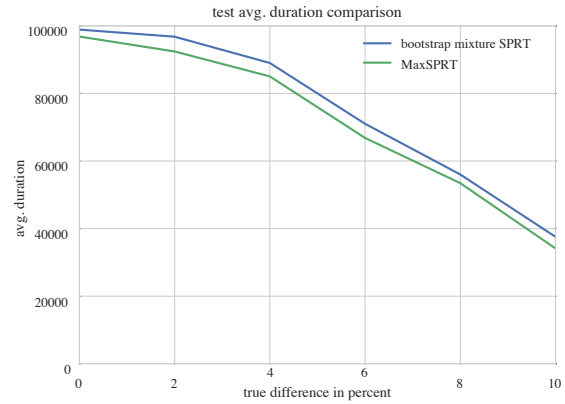


Figure 4: *Test average duration of the bootstrap mixture SPRT for different metrics.*



(a) *Test power for the bootstrap mixture SPRT and MaxSPRT.*



(b) *Average test duration for the bootstrap mixture SPRT and MaxSPRT.*

Figure 5: Comparison of the bootstrap mixture SPRT with MaxSPRT.

data generation distribution is unknown or hard to model, which typically is the case.

7. REFERENCES

[1] T. Crook, B. Frasca, R. Kohavi, and R. Longbotham. Seven pitfalls to avoid when running controlled experiments on the web. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 1105–1114, New York, NY, USA, 2009. ACM.

[2] A. C. Davison, D. V. Hinkley, and B. Worton. Bootstrap likelihoods. *Biometrika*, 79(1):113–130, 1992.

[3] B. Efron. Bayes and likelihood calculations from confidence intervals. *Biometrika*, 80(1):3–26, 1993.

[4] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Chapman & Hall/CRC, 1993.

[5] M. Goodson. Most winning a/b test results are illusory. Whitepaper, Qubit, Jan 2014.

[6] P. Hall. On the bootstrap and likelihood-based confidence regions. *Biometrika*, 74(3):481–493, 1987.

[7] P. Hall. *The Bootstrap and Edgeworth Expansion*. Springer Series in Statistics. Springer, Mar 1997.

[8] R. Johari, L. Pekelis, and D. J. Walsh. Always valid inference: Bringing sequential analysis to a/b testing. http://arxiv.org/abs/1512.04922, Feb 2016.

[9] E. Kharitonov, A. Vorobev, C. Macdonald, P. Serdyukov, and I. Ounis. Sequential testing for early stopping of online experiments. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 473–482, New York, NY, USA, 2015. ACM.

[10] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. Controlled experiments on the web: Survey and practical guide. *Data Min. Knowl. Discov.*, 18(1):140–181, Feb. 2009.

[11] 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.

[12] T. L. Lai. Sequential analysis: Some classical problems and new challenges. *Statistica Sinica*, 11(2):303–351, 2001.

[13] E. Lehmann. *Elements of Large-Sample Theory*, chapter 6, pages 363–450. Springer Texts in Statistics. Springer, Aug 2004.

[14] L. Pekelis, D. Walsh, and R. Johari. The new stats engine. Whitepaper, Optimizely, Jan 2015.

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

[16] M. Pollak. Optimality and almost optimality of mixture stopping rules. *The Annals of Statistics*, 6(4):910–916, 1978.

[17] H. Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.

[18] J. P. Simmons, L. D. Nelson, and U. Simonsohn. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci*, 22(11):1359–1366, Nov 2011.

[19] G. M. van Kempen and L. J. van Vliet. Mean and variance of ratio estimators used in fluorescence ratio imaging. *Cytometry*, 39(4):300–305, Apr 2000.

[20] A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.