S. De and M. Baron. Step-up and step-down methods for testing multiple hypotheses in sequential experiments, *J. of Stat. Planning and Inference*, 142: 2059-2070, 2012.

Remark:

Tables 1 and 2 of this paper compare performance of our proposed procedures with other methods. Results for the non-sequential Holm procedure and the Bartroff–Lai multistage step-down procedure are taken from [38].

It should be noted that the "sample size" of a clinical trial in Table 1 of [38] is actually the total number of sampled components. Optimality of sequential procedures in [38] is also measured in terms of the total number of sampled components.

However, in our paper, the "sample size" is defined as the number of sampled patients. Therefore, our procedures are not to be compared with the Bartroff-Lai rule.

Shyamal and Michael

Step-up and step-down methods for testing multiple hypotheses in sequential experiments

Shyamal K. De and Michael Baron

Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, TX 75080; shyamal@utdallas.edu, mbaron@utdallas.edu.

Abstract

Sequential methods are developed for testing multiple hypotheses, resulting in a statistical decision for each individual test and controlling the familywise error rate and the familywise power in the strong sense. Extending the ideas of step-up and step-down methods for multiple comparisons to sequential designs, the new techniques improve over the Bonferroni and closed testing methods proposed earlier by a substantial reduction of the expected sample size.

Keywords: Bonferroni methods, Familywise error rate, Holm procedure, Sequential probability ratio test, Stopping boundaries, Wald approximation

2000 MSC: 62L, 62F03, 62H15

1. Introduction

1.1. Motivation

The problem of multiple inferences in sequential experiments arises in many fields. Typical applications are in sequential clinical trials with both efficacy and safety endpoints ([1]) or several outcome measures of efficacy ([2], [3]), acceptance sampling with several different criteria of acceptance

([4], [5]), multichannel change-point detection ([6], [7]) and in microarray experiments ([8]). It is often necessary to find the statistical answer to each posed question by testing each individual hypothesis rather than giving one global answer by combining all the tests into one and testing a composite hypothesis.

Methods developed in this article aim to test *multiple hypotheses* based on sequentially collected data, resulting in *individual decisions* for each individual test. They control the familywise error rate and the familywise power in the strong sense. That is, both probabilities of rejecting at least one true null hypothesis and accepting at least one false null hypothesis are kept within the chosen levels α and β under any set of true hypotheses. This condition is a multi-testing analogue of controlling both probabilities of Type I and Type II errors in sequential experiments. As a result, the *familywise power*, defined as the probability of detecting *all* significant differences at the specified alternative parameter values, is controlled at the level $(1 - \beta)$ (see [9] for three alternative definitions of familywise power).

Under these conditions, proposed stopping rules and decision rules achieve substantial reduction of the expected sample size over all the existing (to the best of our knowledge) sequential multiple testing procedures.

1.2. Sequential multiple comparisons in the literature

The concept of multiple comparisons is not new in sequential analysis. Sequential methods exist for inferences about multivariate parameters ([10], sec. 6.8 and 7.5). They are widely used in studies where inferences about individual parameters are not required.

Most of the research in sequential multiple testing is limited to two types of problems.

One type is the study of several ($k > 2$) treatments comparing their effects. Sampled units are randomized to k groups where treatments are administered. Based on the observed responses, one typically tests a composite null hypothesis $H_0 : \theta_1 = \ldots = \theta_k$ against H_A : not H_0, where θ_j is the effect of treatment j for $j = 1, \ldots, k$ ([11], [12], [13], [14], [15], chap. 16, [16], [17], [18], [19], chap. 8). Sometimes each treatment is compared to the accepted standard (e.g., [20]), and often the ultimate goal is selection of the best treatment ([21], [22]).

The other type of studies involves a sequentially observed sequence of data that needs to be classified into one of several available sets of models. In a parametric setting, a null hypothesis $H_0 : \theta \in \Theta_0$ is tested against several alternatives, $H_1 : \theta \in \Theta_1$ vs ... vs $H_k : \theta \in \Theta_k$, where θ is the common parameter of the observed sequence ([23], [24], [25], [26]).

The optimal stopping rules for such tests are (naturally!) extensions of the classical Wald's sequential probability ratio tests ([27], [28], [29], [30]). For the case of three alternative hypotheses, Sobel and Wald ([31]) obtained a set of four stopping boundaries for the likelihood-ratio statistic. Their solution was generalized to a larger number of alternatives resulting in the *multi-hypothesis sequential probability ratio tests* ([32], [33]).

1.3. Our goal - simultaneous testing of individual hypotheses

The focus of this paper is different and more general. We assume that the sequence of sampled units is observed to answer several questions about its parameters. Indeed, once the sampling cost is already spent on each sampled

unit, it is natural to use it to answer more than just one question! Therefore, there are *d individual hypotheses* about parameters $\theta_1, \ldots, \theta_d$ of sequentially observed vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$,

$$\begin{aligned} H_0^{(1)} &: \theta_1 \in \Theta_{01} \quad \text{vs} \quad H_A^{(1)} : \theta_1 \in \Theta_{11}, \\ H_0^{(2)} &: \theta_2 \in \Theta_{02} \quad \text{vs} \quad H_A^{(2)} : \theta_2 \in \Theta_{12}, \\ &\ldots \\ H_0^{(d)} &: \theta_k \in \Theta_{0d} \quad \text{vs} \quad H_A^{(d)} : \theta_k \in \Theta_{1d}. \end{aligned} \tag{1}$$

A few sequential procedures have been proposed for multiple tests similar to (1). One can conduct individual sequential tests of $H_0^{(1)}, \ldots, H_0^{(d)}$ and stop after the first rejection or acceptance, as in [15], chap. 15. Hypotheses that are not rejected at this moment will be accepted, conservatively protecting the familywise Type I error rate (FWER-I).

Alternatively, one can assign level α_j and the corresponding Pocock or O'Brien-Fleming rejection boundary to the jth hypothesis. Then one conducts sequential or group sequential tests in a hierarchical manner, as proposed in [34], [35], and [36] for testing primary, secondary, and possibly tertiary endpoints of a clinical trial. This procedure controls FWER-I at the level $\alpha = \sum \alpha_j$.

A different approach proposed in [37] and further developed in [38] allows to control FWER-I by testing *a closed set* of hypotheses. Along with the individual hypotheses $H_0^{(1)}, \ldots, H_0^{(d)}$, this method requires to test all the composite hypotheses consisting of intersections $\cap H_0^{(j_k)}$, $1 \le j_k \le d$, $1 \le k \le d$. This results in mandatory testing of $(2^d - 1)$ instead of d hypotheses. As shown in Section 4, controlling the overall familywise Type I error rate will then require a rather large expected sample size.

While focusing on the Type I FWER, these procedures do not control the *familywise Type II error rate* and the familywise power. On the other hand, a Type II error, for example, on one of the tests of a safety clinical trial implies a failure to notice a side effect of a treatment, which is important to control.

Notice that sequential tests of single hypotheses are able to control probabilities of both the Type I and Type II errors. Extending this to multiple testing, our goal is to control *both familywise error rates* I and II and to do so at a *low sampling cost* by computing the optimal stopping boundaries and the optimal stopping rule followed by the optimal terminal decisions.

1.4. Approach - extension of non-sequential ideas

To approach this problem, we use the step-up and step-down ideas developed for *non-sequential* multiple comparisons. Detailed overviews of non-sequential methods were given at the NSF-CBMS Conference "New Horizons in Multiple Comparison Procedures" in 2001 ([39]), in a 2008 monograph [40], and at the 2009 Workshop on Modern Multiple Testing by Prof. S. Sarkar.

It was noted that the elementary Bonferroni adjustment for multiple comparisons takes care of the familywise error rate at the expense of power. However, some power can be regained by wisely designed *step-up and step-down methods*, ordering p-values of individual tests, choosing one of the ordered p-values, and proceeding from it into one or the other direction making decisions on the acceptance or rejection of individual null hypotheses ([41], [42], [43], [44], [45], [46], [47]).

Lehmann and Romano ([48]) introduced the *generalized error rate*, which is the probability of making at least r incorrect inferences instead of at least

one. This weaker requirement on the error control allows to regain power in studies with a large number of simultaneous inferences, where Bonferroni-type adjustments result in very low significance levels and a great loss of power. The new concept was quickly developed, and multiple comparison methods controlling the generalized error rate were proposed ([49], [50], [51]).

Fixed-sample studies are able to control either the Type I or the Type II error probabilities, but in general, not both. Conversely, Wald's sequential probability test and subsequent sequential procedures for testing a single hypothesis can be designed to satisfy both the given significance level and the given power. Similarly, sequential testing of multiple hypotheses, considered here, can be set to guarantee a strong control of both familywise Type I and Type II error rates,

$$\begin{aligned} FWER_I &= \max_{\mathcal{T}\neq\varnothing} \boldsymbol{P}\{\text{reject at least one true null hypothesis} \mid \mathcal{T}\} \\ &= \max_{\mathcal{T}\neq\varnothing} \boldsymbol{P}\left\{\bigcup_{j\in\mathcal{T}} \text{reject } H_0^{(j)} \mid \mathcal{T}\right\} \qquad (2) \\ FWER_{II} &= \max_{\mathcal{F}\neq\varnothing} \boldsymbol{P}\{\text{accept at least one false null hypothesis} \mid \mathcal{T}\} \\ &= \max_{\mathcal{F}\neq\varnothing} \boldsymbol{P}\left\{\bigcup_{j\in\bar{\mathcal{T}}} \text{accept } H_0^{(j)} \mid \mathcal{T}\right\} \qquad (3) \end{aligned}$$

where $\mathcal{T} \subset \{1, \ldots, d\}$ is the index set of the true null hypotheses and $\mathcal{F} = \bar{\mathcal{T}}$ is the index set of the false null hypotheses.

In the sequel, the non-sequential Bonferroni, step-down Holm, and step-up Benjamini-Hochberg methods for multiple comparisons are generalized to the sequential setting. Essentially, at every step of the multiple testing scheme, the continue-sampling region is inserted between the acceptance and

rejection boundaries in such a way that the resulting sequential procedure controls the intended error rate. Then, the so-called *intersection stopping rule* is introduced that controls both familywise error rates simultaneously.

2. Sequential Bonferroni method

Let us begin with the rigorous formulation of the problem. Suppose a sequence of independent and identically distributed vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \in \boldsymbol{R}^d$ that are observed as a result of purely sequential or group sequential sampling. Components $(X_{i1}, \ldots, X_{id})$ of the i-th random vector may be dependent, and the j-th component has a marginal density $f_j(\cdot \mid \theta_j)$ with respect to a reference measure $\mu_j(\cdot)$. For every $j = 1, \ldots, d$, measures $\{f_j(\cdot \mid \theta_j), \theta_j \in \Theta_j\}$ are assumed to be mutually absolutely continuous, and the Kullback-Leibler information numbers

$$K(\theta_j, \theta_j') = \mathbf{E}_{\theta_j} \log \left\{ f_j(X_j \mid \theta_j) / f_j(X_j \mid \theta_j') \right\}$$

are strictly positive and finite for $\theta_j \neq \theta_j'$.

Consider a battery of one-sided (right-tail, with a suitable parametrization) tests about parameters $\theta_1, \ldots, \theta_d$,

$$H_0^{(j)} : \theta_j \leq \theta_0^{(j)} \text{ vs. } H_A^{(j)} : \theta_j \geq \theta_1^{(j)}, \ j = 1, \ldots, d, \tag{4}$$

A stopping rule T is to be found, accompanied with decision rules $\delta_j = \delta_j(X_{1j}, \ldots, X_{Tj})$, $j = 1, \ldots, d$, on the acceptance or rejection of each of the null hypotheses $H_0^{(1)}, \ldots, H_0^{(d)}$. This procedure has to control both familywise error rates (2) and (3), i.e., guarantee that

$$FWER_I \leq \alpha \quad \text{and} \quad FWER_{II} \leq \beta \tag{5}$$

for pre-assigned $\alpha, \beta \in (0,1)$.

Technically, it is not difficult to satisfy condition (5). Wald's sequential probability ratio test (SPRT) for the j-th hypothesis controls the probabilities of Type I and Type II errors at the given levels α_j and β_j. Choosing $\alpha_j = \alpha/d$ and $\beta_j = \beta/d$, we immediately obtain (5) by the Bonferroni inequality.

This testing procedure is based on log-likelihood ratios

$$\Lambda_n^{(j)} = \sum_{i=1}^{n} \log \frac{f_j(X_{ij} \mid \theta_1^{(j)})}{f_j(X_{ij} \mid \theta_0^{(j)})}, \quad j = 1, \ldots, d, \ n = 1, 2, \ldots.$$

Wald's classical stopping boundaries are

$$a_j = \log \frac{1-\beta_j}{\alpha_j} \quad \text{and} \quad b_j = \log \frac{\beta_j}{1-\alpha_j}. \tag{6}$$

Wald's SPRT for the single j-th hypothesis $H_0^{(j)}$ rejects it (i.e., chooses $H_A^{(j)}$) after n observations if $\Lambda_n^{(j)} \geq a_j$, accepts it (i.e., chooses $H_0^{(j)}$) if $\Lambda_n^{(j)} \leq b_j$, and continues sampling if $\Lambda_n^{(j)} \in (b_j, a_j)$.

Assuming that the marginal distributions of $\Lambda_1^{(j)}$ have the monotone likelihood ratio property (e.g. [52]), the error probabilities are maximized when $\theta_j = \theta_0^{(j)}$ and when $\theta_j = \theta_1^{(j)}$, respectively, for all $j = 1, \ldots, d$. Then, separately performed SPRT for the j-th hypothesis with stopping boundaries (6) controls the probabilities of Type I and Type II errors *approximately*,

$$\begin{aligned} \boldsymbol{P}\left\{\Lambda_{T_j}^{(j)} \geq a_j \mid \theta_j = \theta_0^{(j)}\right\} &\approx \alpha_j, \\ \boldsymbol{P}\left\{\Lambda_{T_j}^{(j)} \leq b_j \mid \theta_j = \theta_1^{(j)}\right\} &\approx \beta_j, \end{aligned} \tag{7}$$

where $T_j = \inf\left\{n : \ \Lambda_n^{(j)} \notin (b_j, a_j)\right\}$ ([53], [27], [28]). This *Wald's approximation* ([54]) results from ignoring the overshoot over the stopping boundary

and assuming that having just crossed the stopping boundary for the first time, the log-likelihood ratio $\Lambda_n^{(j)}$ approximately equals to that boundary.

Extending SPRT to the case of multiple hypothesis, $d \geq 2$, continue sampling until all the d tests reach decisions. Define the stopping rule

$$T = \inf \left\{ n \ : \ \bigcap_{j=1}^{d} \left\{ \Lambda_n^{(j)} \notin (b_j, a_j) \right\} \right\}. \tag{8}$$

Lemma 1. *For any pairs (a_j, b_j), the stopping rule defined by (8) is proper.*

Proof. Section 5.1. □

Accepting or rejecting the j-th null hypothesis at time T depending on whether $\Lambda_T^{(j)} \leq b_j$ or $\Lambda_T^{(j)} \geq a_j$, we could obtain (approximately, subject to Wald's approximation) a strong control of probabilities of Type I and Type II errors by the Bonferroni inequality,

$$\begin{aligned} \boldsymbol{P}\left\{\text{at least one Type I error}\right\} &\leq \sum_{j=1}^{d} \boldsymbol{P}\left\{\Lambda_T^{(j)} \geq a_j\right\} \leq \sum_{j=1}^{d} \alpha_j = \alpha; \\ \boldsymbol{P}\left\{\text{at least one Type II error}\right\} &\leq \sum_{j=1}^{d} \boldsymbol{P}\left\{\Lambda_T^{(j)} \leq b_j\right\} \leq \sum_{j=1}^{d} \beta_j = \beta. \end{aligned} \tag{9}$$

However, Wald's approximation is only accurate when the overshoot of Λ_T over the stopping boundary is negligible. When testing d hypotheses, the corresponding d log-likelihood ratios may cross their respective boundaries at different times. Then, at the stopping time T, when sampling is halted, a number of log-likelihood ratios may be deep inside the stopping region, creating a considerable overshoot. Wald's approximation is no longer accurate for the stopping time T! It has to be replaced by rigorous statements.

Lemma 2. *Let T be a proper stopping time with respect to the vector sequence $(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots)$, such that*

$$\boldsymbol{P}\left\{\Lambda_T^{(j)} \in (b, a) \mid \mathcal{T}\right\} = 0$$

for some $j \in \{1, \ldots, d\}$, $b < 0 < a$, and any combination of true null hypotheses $\mathcal{T}$. Consider a test that rejects $H_0^{(j)}$ at time T if and only if $\Lambda_T^{(j)} \geq a$. For this test,

$$\boldsymbol{P}\left\{\textit{Type I error on } H_0^{(j)}\right\} \;\leq\; \boldsymbol{P}\left\{\Lambda_T^{(j)} \geq a \mid \theta_0^{(j)}\right\} \leq e^{-a}, \tag{10}$$

$$\boldsymbol{P}\left\{\textit{Type II error on } H_0^{(j)}\right\} \;\leq\; \boldsymbol{P}\left\{\Lambda_T^{(j)} \leq b \mid \theta_1^{(j)}\right\} \leq e^{b}. \tag{11}$$

Proof. Section 5.2. □

Avoiding the use of Wald's (and any other) approximation, replace Wald's stopping boundaries (6) for $\Lambda_n^{(j)}$ by

$$a_j = -\log \alpha_j \quad \text{and} \quad b_j = \log \beta_j \tag{12}$$

and use the stopping rule (8). Then, according to Lemma 2, the corresponding test of $H_0^{(j)}$ controls the Type I and Type II error probabilities rigorously at levels $e^{-a_j} = \alpha_j$ and $e^{b_j} = \beta_j$. Therefore, by the Bonferroni arguments in (9), the described multiple testing procedure controls both error rates. The following theorem is then proved.

Theorem 1. *Sequential Bonferroni procedure for testing multiple hypotheses (4) with the stopping rule (8), rejection regions $\Lambda_T^{(j)} \geq -\log(\alpha/d)$, and acceptance regions $\Lambda_T^{(j)} \leq \log(\beta/d)$ controls both error rates at levels $FWER_I \leq \alpha$ and $FWER_{II} \leq \beta$.*

Further development of the Bonferroni methods and comparison of the associated stopping rules can be found in [55].

3. Step-down and step-up methods

Since Bonferroni methods are based on an inequality that appears rather crude for moderate to large d, controlling both familywise rates I and II can only be done at the expense of a large sample size. For non-sequential statistical inferences, a number of elegant stepwise (step-up and step-down) methods have been proposed, attaining the desired FWER-I and improving over the Bonferroni methods in terms of the required sample size. In this Section, we develop a similar approach for sequential experiments.

Following the Holm method for multiple comparisons (e.g., [56]), we order the tested hypotheses $H_0^{(1)}, \ldots, H_0^{(d)}$ according to the significance of the collected evidence against them and set the significance levels for individual tests to be

$$\alpha_1 = \frac{\alpha}{d},\ \alpha_2 = \frac{\alpha}{d-1},\ \alpha_3 = \frac{\alpha}{d-2},\ \ldots,\ \alpha_j = \frac{\alpha}{d+1-j},\ \ldots,\ \alpha_d = \alpha.$$

Similarly, in order to control the familywise Type II error rate, choose

$$\beta_j = \frac{\beta}{d+1-j} \quad \text{for } j = 1, \ldots, d.$$

Comparing with Bonferroni method, increasing the individual Type I and Type II error probabilities has to cause the familywise error rates to increase. On the other hand, since all the stopping boundaries become *tighter*, this will necessarily reduce the expected sample size $\mathbf{E}(T)$ under any combination of true and false hypotheses, $\mathcal{T}$ and $\mathcal{F}$. Therefore, if both rates can still be controlled at the pre-specified levels α and β, then the resulting stepwise procedure is an improvement of Bonferroni schemes under the given constraints on FWE rates.

In the following algorithm, we combine the stepwise idea for efficient multiple testing with Wald's sequential probability ratio testing of individual hypotheses. The first scheme controls FWER-I, the second scheme controls FWER-II, and the third "intersection" scheme based on them controls *both* familywise error rates.

Scheme 1: Reject one by one, accept all (step-down scheme). Choose the stopping boundaries $a_j = -\log \alpha_j = -\log(\alpha/(d-j+1))$ and arbitrary $b_j < 0$ for $j = 1, \ldots, d$.

After observing vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, order the log-likelihood ratio statistics in their *non-increasing* order,

$$\Lambda_n^{[1]} \geq \Lambda_n^{[2]} \geq \ldots \geq \Lambda_n^{[d]},$$

and let $H_0^{[j]}$ for $j = 1, \ldots, d$ be the corresponding tested null hypotheses arranged in the same order. Proceed according to a *step-down* scheme, from the most significant log-likelihood ratio statistic $\Lambda_n^{[1]}$ down to the second most significant statistic, etc.

Step 1.	If $\Lambda_n^{[1]}$	$\leq$	b_1	then	accept all $H_0^{[1]} \ldots, H_0^{[d]}$
	If $\Lambda_n^{[1]}$	$\in$	(b_1, a_1)	then	continue sampling; collect $\boldsymbol{X}_{n+1}$
	If $\Lambda_n^{[1]}$	$\geq$	a_1	then	reject $H_0^{[1]}$ and go to $\Lambda_n^{[2]}$
Step 2.	If $\Lambda_n^{[2]}$	$\leq$	b_2	then	accept all $H_0^{[2]} \ldots, H_0^{[d]}$
	If $\Lambda_n^{[2]}$	$\in$	(b_2, a_2)	then	continue sampling; collect $\boldsymbol{X}_{n+1}$
	If $\Lambda_n^{[2]}$	$\geq$	a_2	then	reject $H_0^{[2]}$ and go to $\Lambda_n^{[3]}$

etc. (Fig. 1).

Sampling continues while at least one ordered log-likelihood ratio $\Lambda_n^{[j]}$

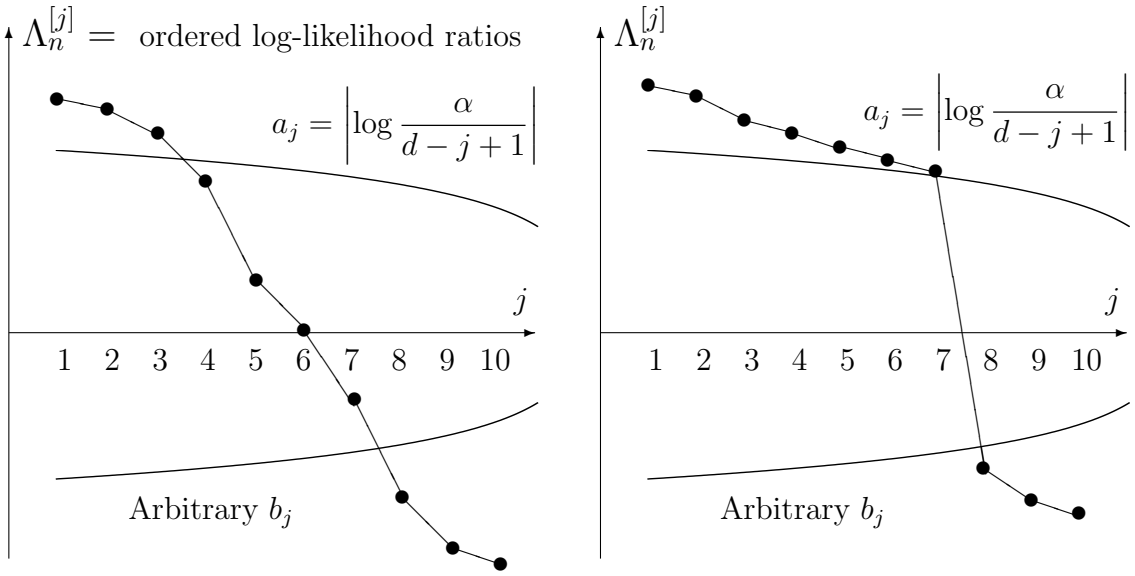

Figure 1: Example of Scheme 1. On the left, sampling continues. On the right, stop sampling; reject $H_0^{[1]}, \ldots, H_0^{[7]}$; accept $H_0^{[8]}, \ldots, H_0^{[10]}$.

belongs to its continue-sampling region (b_j, a_j). The stopping rule corresponding to this scheme is

$$T_1 = \inf\left\{ n \ : \ \bigcap_{j=1}^{d} \Lambda_n^{[j]} \notin (b_j, a_j) \right\}. \tag{13}$$

Theorem 2. *The stopping rule T_1 is proper; Scheme 1 strongly controls the familywise Type I error rate. That is, for any set $\mathcal{T}$ of true hypotheses,*

$$\boldsymbol{P}_{\mathcal{T}}\{T_1 < \infty\} = 1$$

and

$$\boldsymbol{P}_{\mathcal{T}}\{\textit{at least one Type I error}\} \leq \alpha. \tag{14}$$

Proof. Section 5.3. □

Theorem 2 holds for arbitrary $b_j < 0$, thus it also shows that the rejection boundary a_j alone controls FWER-I.

Symmetrically to Scheme 1, introduce the following Scheme 2 which controls FWER-II through the choice of acceptance boundary b_j alone.

Scheme 2: Accept one by one, reject all (step-up scheme). For this scheme, choose the stopping boundaries $b_j = \log \beta_j = \log(\beta/(d-j+1))$ and arbitrary $a_j > 0$ for $j = 1, \ldots, d$.

After observing $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, order the log-likelihood ratio statistics in their *non-decreasing* order,

$$\Lambda_n^{\{1\}} \le \Lambda_n^{\{2\}} \le \ldots \le \Lambda_n^{\{d\}},$$

and let $H_0^{\{j\}}$ for $j = 1, \ldots, d$ be the corresponding tested null hypotheses arranged in the same order. Proceed according to the *step-up scheme*, from the least significant log-likelihood ratio to the second least significant, etc.

Step 1.	If $\Lambda_n^{\{1\}}$	$\ge$	a_1	then	reject all $H_0^{\{1\}} \ldots, H_0^{\{d\}}$
	If $\Lambda_n^{\{1\}}$	$\in$	(b_1, a_1)	then	continue sampling; collect $\boldsymbol{X}_{n+1}$
	If $\Lambda_n^{\{1\}}$	$\le$	b_1	then	accept $H_0^{\{1\}}$ and go to $\Lambda_n^{\{2\}}$
Step 2.	If $\Lambda_n^{\{2\}}$	$\ge$	a_2	then	reject all $H_0^{\{2\}} \ldots, H_0^{\{d\}}$
	If $\Lambda_n^{\{2\}}$	$\in$	(b_2, a_2)	then	continue sampling; collect $\boldsymbol{X}_{n+1}$
	If $\Lambda_n^{\{2\}}$	$\le$	b_2	then	accept $H_0^{\{2\}}$ and go to $\Lambda_n^{\{3\}}$

etc.

According to this scheme, the stopping rule is, similarly to (13),

$$T_2 = \inf \left\{ n \ : \ \bigcap_{j=1}^{d} \Lambda_n^{\{j\}} \notin (b_j, a_j) \right\}, \tag{15}$$

Theorem 3. *The stopping rule T_2 is proper; Scheme 2 strongly controls the familywise Type II error rate. That is, for set $\mathcal{T}$ of true hypotheses,*

$$\boldsymbol{P}_{\mathcal{T}} \{T_2 < \infty\} = 1$$

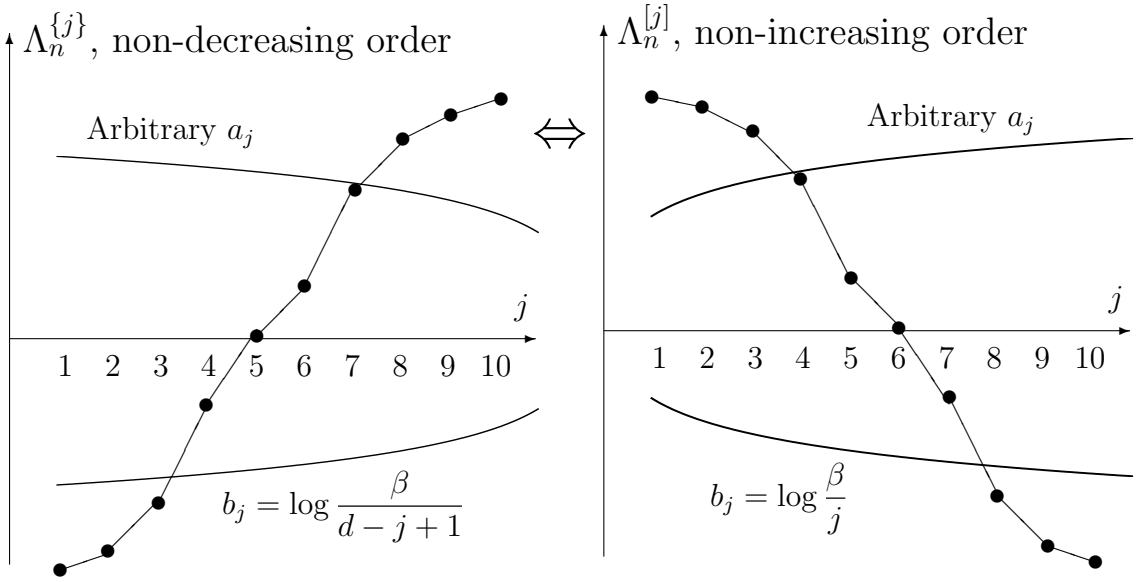

Figure 2: Scheme 2 in terms of log-likelihood ratios in the non-decreasing order (left), and equivalently, in the non-increasing order (right).

and

$$\boldsymbol{P}_{\mathcal{T}}\{\textit{at least one Type II error}\} \leq \beta. \tag{16}$$

Proof. Section 5.4. □

Notice that Scheme 2 can be equivalently formulated in terms of the non-increasing statistics $\Lambda_n^{[j]}$ instead of non-decreasing $\Lambda_n^{\{j\}}$. Indeed, one rearranges the log-likelihood ratios from their non-increasing ordering to their nonincreasing ordering by simply reverting the order, i.e.,

$$\Lambda_n^{\{j\}} = \Lambda_n^{[d-j+1]}.$$

Rearranging the boundary values accordingly, i.e., replacing $b_j = \log(\beta/d - j + 1)$ with $b_j = \log(\beta/j)$, one obtains Scheme 2 in terms of $\Lambda_n^{[j]}$ instead of $\Lambda_n^{\{j\}}$. In short,

$$\Lambda_n^{\{j\}} \leq \log \frac{\beta}{d-j+1} \quad \Leftrightarrow \quad \Lambda_n^{\{d-j+1\}} \leq \log \frac{\beta}{j} \quad \Leftrightarrow \quad \Lambda_n^{[j]} \leq \log \frac{\beta}{j}.$$

This is illustrated in Fig. 2.

Comparing the logic of Schemes 1 and 2, we see that the step-down Scheme 1 starts with the most significant log-likelihood ratio statistic $\Lambda_n^{[1]}$ and carefully/conservatively rejects one null hypothesis at a time. It focuses on controlling Type I errors of wrong rejection and results in controlling the overall FWER-I.

On the contrary, the step-up Scheme 2 starts with the least significant statistic and conservatively accepts one null at a time, controlling FWER-II.

It is actually possible to combine both schemes and to develop a sequential procedure controlling both familywise error rates as follows.

The Intersection Scheme. Combining Schemes 1 and 2, define the *intersection stopping rule*

$$T^* = \inf\left\{ n \,:\, \bigcap_{j=1}^{d} \Lambda_n^{[j]} \notin (b_j^*, a_j^*) \right\},\ a_j^* = -\log\frac{\alpha}{d-j+1},\ b_j^* = \log\frac{\beta}{j}. \quad (17)$$

Acceptance and rejection regions in this case are simply *intersections* of rejection and acceptance regions for $\Lambda_n^{[j]}$ for Schemes 1 and 2, as long as the lower boundary in Scheme 1 and the upper boundary in Scheme 2 are inside the interval $[\log\beta, -\log\alpha]$.

According to the Intersection Scheme, reject the hypothesis $H_0^{[j]}$ corresponding to the j-th ordered log-likelihood ratio $\Lambda_n^{[j]}$ if $\Lambda_{T^*}^{[j]} \geq a_j^*$ and accept it if $\Lambda_{T^*}^{[j]} \leq b_j^*$ (Fig. 3).

Being a special case of Scheme 1 and Scheme 2 at the same time, the Intersection Scheme controls both FWER-I and FWER-II.

Theorem 4. *The stopping rule T^* is proper; the Intersection Scheme strongly controls both familywise Type I and Type II error rates. That is, for any set*

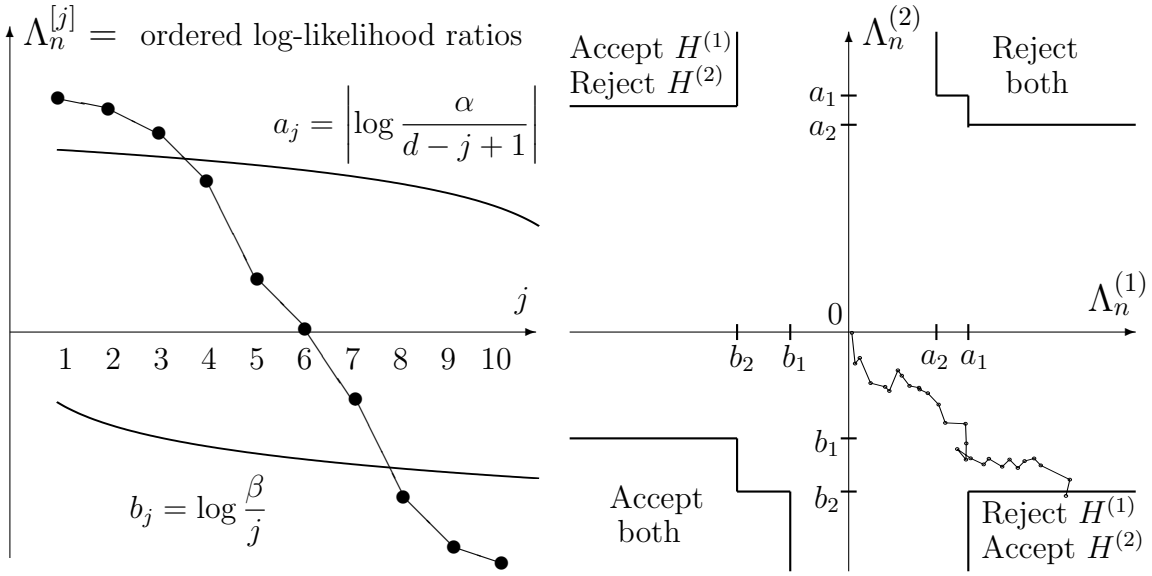

Figure 3: The Intersection Scheme. Stopping boundaries for the ordered log-likelihood ratios for $d = 10$ tests, snap shot after n observations (left). Stopping boundaries for the unordered log-likelihood ratios for $d = 2$ tests, the entire path until the stopping time T^* (right).

$\mathcal{T}$ of true hypotheses,

$$\boldsymbol{P}_{\mathcal{T}}\{T^* < \infty\} = 1,$$

$$\boldsymbol{P}_{\mathcal{T}}\{\textit{at least one Type I error}\} \le \alpha,$$

$$\boldsymbol{P}_{\mathcal{T}}\{\textit{at least one Type II error}\} \le \beta.$$

Proof. Section 5.5. □

4. Comparison of multiple testing schemes

Performance of the proposed schemes is evaluated and compared with the Holm-Bonferroni non-sequential multiple testing procedure and the multistage step-down procedure of [38].

First, consider testing three null hypotheses,

$$\begin{aligned} H_0^{(1)} : \theta_1 = 0 \quad &\text{vs} \quad H_A^{(1)} : \theta_1 = 0.5, \\ H_0^{(2)} : \theta_2 = 0 \quad &\text{vs} \quad H_A^{(2)} : \theta_2 = 0.5, \\ H_0^{(3)} : \theta_3 = 0.5 \quad &\text{vs} \quad H_A^{(3)} : \theta_3 = 0.75, \end{aligned}$$

based on a sequence of random vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$, where $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3})$, $X_{i1} \sim Normal(\theta_1, 1)$, $X_{i2} \sim Normal(\theta_2, 1)$, and $X_{i3} \sim Bernoulli(\theta_3)$, which is the scenario considered in [38].

For each combination of null hypotheses and alternatives, $N = 55,000$ random sequences are simulated (omitting the combinations $H_A^{(1)} \cap H_0^{(2)} \cap H_0^{(3)}$ and $H_A^{(1)} \cap H_0^{(2)} \cap H_A^{(3)}$, because of their interchangeability with $H_0^{(1)} \cap H_A^{(2)} \cap H_0^{(3)}$ and $H_0^{(1)} \cap H_A^{(2)} \cap H_A^{(3)}$, respectively, therefore yielding exactly the same performance of the considered stopping and decision rules). Based on them, the familywise Type I and Type II error rates are estimated along with the expected stopping time and the standard error of our estimator of the expected stopping time. Scheme 1 and the Intersection Scheme are set to control $FWER_I \leq 0.05$. Also, Scheme 2 and the Intersection Scheme are set to control $FWER_{II} \leq 0.10$.

Results are shown in Tables 1-2. It can be seen that under each combination of true null hypotheses, the expected sample sizes of sequential Bonferroni, step-down Scheme 1, step-up Scheme 2, and the Intersection Scheme are of the same order; however, they are about a half of the expected size of the closed testing procedure or the fixed-sample size required by the non-sequential Holm-Bonferroni procedure.

Advantage of stepwise schemes versus the sequential Bonferroni scheme

Procedure	Expected st. time	Standard Error	FWER-I	FWER-II
Under $H_0^{(1)}$, $H_0^{(2)}$, $H_0^{(3)}$; $\theta_1 = 0$, $\theta_2 = 0$, $\theta = 0.5$				
Bonferroni	46.8	0.10	0.022	–
Scheme 1	35.4	0.09	0.026	–
Scheme 2	35.7	0.08	0.055	–
Intersection Scheme	37.0	0.09	0.023	–
Nonsequential Holm	105.0	–	0.040	–
Multistage step-down	104.6	–	0.048	–
Under $H_0^{(1)}$, $H_0^{(2)}$, $H_A^{(3)}$; $\theta_1 = 0$, $\theta_2 = 0$, $\theta = 0.75$				
Bonferroni	49.3	0.10	0.013	0.010
Scheme 1	41.2	0.09	0.021	0.034
Scheme 2	41.3	0.09	0.036	0.027
Intersection Scheme	45.7	0.09	0.019	0.028
Nonsequential Holm	105.0	–	0.044	–
Multistage step-down	98.3	–	0.042	–
Under $H_0^{(1)}$, $H_A^{(2)}$, $H_0^{(3)}$; $\theta_1 = 0$, $\theta_2 = 0.65$, $\theta = 0.5$				
Bonferroni	41.8	0.09	0.017	0.001
Scheme 1	33.9	0.07	0.026	0.005
Scheme 2	36.1	0.08	0.043	0.004
Intersection Scheme	38.7	0.08	0.024	0.004
Nonsequential Holm	105.0	–	0.044	–
Multistage step-down	96.9	–	0.049	–

Table 1: Comparison of the proposed sequential schemes with nonsequential Holm and sequential closed testing procedures.

Procedure	Expected st. time	Standard Error	FWER-I	FWER-II
Under $H_0^{(1)}$, $H_A^{(2)}$, $H_A^{(3)}$; $\theta_1 = 0$, $\theta_2 = 0.5$, $\theta = 0.75$				
Bonferroni	51.6	0.10	0.006	0.021
Scheme 1	43.8	0.09	0.016	0.055
Scheme 2	43.1	0.09	0.019	0.034
Intersection Scheme	48.0	0.10	0.016	0.030
Nonsequential Holm	105.0	–	0.043	–
Multistage step-down	92.3	–	0.032	–
Under $H_A^{(1)}$, $H_A^{(2)}$, $H_0^{(3)}$; $\theta_1 = 0.5$, $\theta_2 = 0.5$, $\theta = 0.5$				
Bonferroni	51.6	0.10	0.007	0.022
Scheme 1	44.3	0.09	0.016	0.059
Scheme 2	42.3	0.09	0.020	0.037
Intersection Scheme	47.8	0.10	0.016	0.032
Nonsequential Holm	105.0	–	0.038	–
Multistage step-down	93.1	–	0.027	–
Under $H_A^{(1)}$, $H_A^{(2)}$, $H_A^{(3)}$; $\theta_1 = 0.5$, $\theta_2 = 0.5$, $\theta = 0.75$				
Bonferroni	53.7	0.10	–	0.029
Scheme 1	42.1	0.08	–	0.073
Scheme 2	42.2	0.09	–	0.038
Intersection Scheme	43.9	0.09	–	0.031
Nonsequential Holm	105.0	–	–	–
Multistage step-down	86.1	–	–	–

Table 2: Comparison of the proposed sequential schemes with nonsequential Holm and sequential closed testing procedures (continued).

is more significant for a larger number of tests. This is seen in Table 3 for different multiple testing problems. Reduction in the expected sample size ranges from 6% when 50% of null hypotheses are true to 16.5% when most null hypotheses are either true or false. Results are also based on $N = 55,000$ simulated sequences for each considered scenario.

The last example in Table 3 deals with correlated components of the observed random vectors. Indeed, all the results in this article make no assumption about the joint distribution of $(X_{i1}, \ldots, X_{id})$ for each $i = 1, 2, \ldots$. When components are correlated, positively or negatively, the expected sample size of each procedure should reduce.

We also notice that results of Theorems 1-4 are based on Bonferroni-type inequalities and corollaries from them. For a large number of tests, Bonferroni inequality tends to be rather crude, and therefore, the familywise error rates, guaranteed by Theorems 1-4, are often satisfied with a rather wide margin. This certainly leaves room for the improvement of the proposed sequential testing schemes!

5. Appendix: Proofs

5.1. Proof of Lemma 1

By the weak law of large numbers,

$$\boldsymbol{P}\left\{\Lambda_n^{(j)} \in (b_j, a_j) \mid H_0^{(j)}\right\} \to 0 \text{ and } \boldsymbol{P}\left\{\Lambda_n^{(j)} \in (b_j, a_j) \mid H_A^{(j)}\right\} \to 0, \text{ as } n \to \infty,$$

because the non-zero expected values of $\Lambda_n^{(j)}$ are guaranteed by the assumptions of Section 2 on Kullback-Leibler information numbers. Then

$$\boldsymbol{P}_{\mathcal{T}}\{T > n\} \leq \sum_{j=1}^{d} \boldsymbol{P}_{\mathcal{T}}\left\{\Lambda_n^{(j)} \in (b_j, a_j)\right\} \to 0,$$

Procedure	Expected st. time	Standard Error	FWER-I	FWER-II
Scenario: $\mu_1, \ldots, \mu_{10}$ are parameters of Normal$(\cdot, 1)$ distributions of $X_{i,1}, \ldots, X_{i,10}$; $p_1, \ldots, p_{10}$ are parameters of Bernoulli$(\cdot)$ distributions of $X_{i,11}, \ldots, X_{i,20}$; test $\mu_j = 0$ vs 0.5 and $p_j = 0.5$ vs 0.75; odd-numbered null hypotheses are true				
Bonferroni	118.4	0.13	0.002	0.003
Intersection Scheme	111.1	0.13	0.004	0.006
Scenario: $\mu_1, \ldots, \mu_{10}$ are parameters of Normal$(\cdot, 1)$ distributions of $X_{i,1}, \ldots, X_{i,10}$; $p_1, \ldots, p_{10}$ are parameters of Bernoulli$(\cdot)$ distributions of $X_{i,11}, \ldots, X_{i,20}$; test $\mu_j = 0$ vs 0.5 and $p_j = 0.5$ vs 0.75; null hypotheses $H_0^{(1-9,\ 11-19)}$ are true				
Bonferroni	114.6	0.13	0.004	0.001
Intersection Scheme	96.1	0.12	0.007	0.006
Scenario: $\mu_1, \ldots, \mu_{10}$ are parameters of Normal$(\cdot, 1)$ distributions of $X_{i,1}, \ldots, X_{i,10}$; $p_1, \ldots, p_{10}$ are parameters of Bernoulli$(\cdot)$ distributions of $X_{i,11}, \ldots, X_{i,20}$; test $\mu_j = 0$ vs 0.5 and $p_j = 0.5$ vs 0.75; null hypotheses $H_0^{(10,20)}$ are true				
Bonferroni	121.6	0.13	0.001	0.005
Intersection Scheme	101.5	0.12	0.004	0.008
Scenario: $\mu_1, \ldots, \mu_{10}$ are parameters of Poisson$(\cdot)$ distributions of $X_{i,1}, \ldots, X_{i,10}$; $p_1, \ldots, p_{10}$ are parameters of Bernoulli$(\cdot)$ distributions of $X_{i,11}, \ldots, X_{i,20}$; test $\mu_j = 5$ vs 6 and $p_j = 0.5$ vs 0.75; odd-numbered null hypotheses are true				
Bonferroni	148.4	0.18	0.002	0.003
Intersection Scheme	138.7	0.17	0.003	0.006
Scenario: $\mu_1, \ldots, \mu_5$ are parameters of Normal$(\cdot, 1)$ distributions of $X_{i,1}, \ldots, X_{i,5}$; with $Cov(X_{i,j}, X_{i,j'}) = 0.5$ for $\forall j \neq j'$; $p_1, \ldots, p_5$ are parameters of Bernoulli$(\cdot)$ distributions of $X_{i,6}, \ldots, X_{i,10}$; test $\mu_j = 0$ vs 0.5 and $p_j = 0.5$ vs 0.75; odd-numbered null hypotheses are true				
Bonferroni	91.2	0.12	0.002	0.004
Intersection Scheme	85.3	0.12	0.004	0.007

Table 3: Sequential Bonferroni scheme versus sequential stepwise procedures.

and $\boldsymbol{P}_{\mathcal{T}}\{T=\infty\} \le \boldsymbol{P}_{\mathcal{T}}\{T>n\} \to 0$, therefore, $\boldsymbol{P}_{\mathcal{T}}\{T=\infty\}=0$.

5.2. Proof of Lemma 2

The proof is based on Doob's maximal inequality for submartingales (e.g., [57], Sect. 14.6; [58], Sect. 4.5).

For any $j=1,\ldots,d$ and $n=1,2,\ldots$, the likelihood ratio

$$\lambda_n^{(j)} = \exp\left\{\Lambda_n^{(j)}\right\} = \prod_{i=1}^{n} \frac{f_j(X_{ij} \mid \theta_1^{(j)})}{f_j(X_{ij} \mid \theta_0^{(j)})}$$

is a non-negative martingale under $\theta_0^{(j)}$ with respect to the filtration generated by $(X_{1j}, X_{2j}, \ldots)$. Then, by Doob's inequality, for all $N \ge 1$,

$$\boldsymbol{P}\left\{\max_{1\le n\le N} \lambda_n^{(j)} \ge e^a \mid \theta_0^{(j)}\right\} \le e^{-a}\,\mathbf{E}\left\{\lambda_N^{(j)} \mid \theta_0^{(j)}\right\} = e^{-a},$$

and

$$\begin{aligned}
\boldsymbol{P}\left\{\text{Type I error on } H_0^{(j)}\right\} &\le \boldsymbol{P}\left\{\Lambda_T^{(j)} \ge a \mid \theta_0^{(j)}\right\} \\
&\le \boldsymbol{P}\left\{\max_{1\le n\le N} \Lambda_n^{(j)} \ge a \mid \theta_0^{(j)}\right\} + \boldsymbol{P}\left\{T>N \mid \theta_0^{(j)}\right\} \\
&\le e^{-a} + \boldsymbol{P}\left\{T>N \mid \theta_0^{(j)}\right\}. \qquad (18)
\end{aligned}$$

Taking the limit as $N\to\infty$ proves inequality (10) because $\boldsymbol{P}\left\{T>N \mid \theta_0^{(j)}\right\} \to 0$ since the stopping time T is proper.

To prove inequality (11), we notice that $1/\lambda_n^{(j)} = \exp\left\{-\Lambda_n^{(j)}\right\}$ is a non-negative martingale under $\theta_1^{(j)}$. Applying Doob's inequality, we obtain

$$\boldsymbol{P}\left\{\min_{1\le n\le N} \Lambda_n^{(j)} \le b \mid \theta_1^{(j)}\right\} = \boldsymbol{P}\left\{\max_{1\le n\le N} 1/\lambda_n^{(j)} \ge e^{-b} \mid \theta_1^{(j)}\right\} \le e^b,$$

and the arguments similar to (18) conclude the proof.

5.3. Proof of Theorem 2

1. The stopping time T_1 is almost surely bounded by

$$T' = \inf\left\{ n \ : \ \bigcap_{j=1}^{d} \left\{ \Lambda_n^{(j)} \notin (\min b_j, \max a_j) \right\} \right\}. \tag{19}$$

Since T' is proper by Lemma 1, so is T_1.

2. The proof of control of FWER-I borrows ideas from the classical derivation of the experimentwise error rate of the non-sequential Holm procedure, extending the arguments to sequential tests.

Let $\mathcal{T} \subset \{1, \ldots, d\}$ be the index set of true null hypotheses, and $\mathcal{F} = \bar{\mathcal{T}}$ be the index set of the false null hypotheses, with cardinalities $|\mathcal{T}|$ and $|\mathcal{F}|$. Then, arrange the log-likelihood ratios at the stopping time T_1 in their non-increasing order, $\Lambda_{T_1}^{[1]} \geq \ldots \geq \Lambda_{T_1}^{[d]}$ and let m be the smallest index of the ordered log-likelihood ratio that corresponds to a true hypothesis. In other words, if $H_0^{[j]}$ denotes the null hypothesis that is being tested by the log-likelihood ratio $\Lambda_{T_1}^{[j]}$ for $j = 1, \ldots, d$, then m is such that all $H_0^{[j]}$ are false for $j < m$ whereas $H_0^{[m]}$ is true. Thus, there are at least $(m-1)$ false hypotheses, so that $m - 1 \leq |\mathcal{F}| = d - |\mathcal{T}|$.

No Type I error can be made on false hypotheses $H_0^{[1]}, \ldots, H_0^{[m-1]}$. If the Type I error is not made on $H_0^{[m]}$ either, then there is no Type I error at all because according to Scheme 1, acceptance of $H_0^{[m]}$ implies automatic acceptance of the remaining hypotheses $H_0^{[m+1]}, \ldots, H_0^{[d]}$.

Therefore,

$$\begin{aligned} \boldsymbol{P}_{\mathcal{T}}\{\text{at least one Type I error}\} &= \boldsymbol{P}_{\mathcal{T}}\left\{ \Lambda_{T_1}^{[m]} \geq a_m \right\} \leq \boldsymbol{P}_{\mathcal{T}}\left\{ \Lambda_{T_1}^{[m]} \geq a_{d-|\mathcal{T}|+1} \right\} \\ &= \boldsymbol{P}_{\mathcal{T}}\left\{ \max_{j\in\mathcal{T}} \Lambda_{T_1}^{(j)} \geq a_{d-|\mathcal{T}|+1} \right\} \leq \sum_{j\in\mathcal{T}} \boldsymbol{P}_{\mathcal{T}}\left\{ \Lambda_{T_1}^{(j)} \geq a_{d-|\mathcal{T}|+1} \right\}. \end{aligned}$$

Recall that rejection boundaries for Scheme 1 are chosen as $a_j = -\log \alpha_j$, where $\alpha_j = \alpha/(d+1-j)$. Therefore, $a_{d-|\mathcal{T}|+1} = -\log(\alpha/|\mathcal{T}|)$, and by Lemma 2,

$$\boldsymbol{P}_{H_0^{(j)}} \left\{ \Lambda_{T_1}^{(j)} \geq a_{d-|\mathcal{T}|+1} \right\} \leq \exp\left\{ -a_{d-|\mathcal{T}|+1} \right\} = \alpha/|\mathcal{T}|.$$

Finally, we have

$$\boldsymbol{P}_{\mathcal{T}} \{\text{at least one Type I error}\} \leq \sum_{j \in \mathcal{T}} \alpha/|\mathcal{T}| = \alpha.$$

5.4. Main steps of the proof of Theorem 3

Ideas of Section 5.3 are now translated to Scheme 2 and control of FWER-II. Let us outline the main steps of the proof, especially because control of FWER-II has not been studied in sequential multiple testing, to the best of our knowledge.

1. Similarly to T_1, the stopping time T_2 is also bounded by the proper stopping rule (19), and therefore, it is also proper.

2. Following Scheme 2, arrange $\Lambda_{T_2}^{(j)}$ in their non-decreasing order, $\Lambda_{T_2}^{\{1\}} \leq \ldots \leq \Lambda_{T_2}^{\{d\}}$. Then let ℓ be the smallest index of the ordered log-likelihood ratio that corresponds to a false null hypothesis, so that all $H_0^{\{1\}}, \ldots, H_0^{\{\ell-1\}}$, corresponding to $\Lambda_{T_2}^{\{1\}}, \ldots, \Lambda_{T_2}^{\{\ell-1\}}$, are true but $H_0^{\{\ell\}}$ is false. The number of true hypotheses is then at least $(\ell-1)$, so that $\ell \leq |\mathcal{T}| + 1 = d - |\mathcal{F}| + 1$, where $|\mathcal{T}|$ and $|\mathcal{F}|$ are the numbers of true and false null hypotheses.

If any Type II error is made during Scheme 2, then it has to occur on $H_0^{\{\ell\}}$, because its (correct) rejection leads to the automatic (correct) rejection of the remaining hypotheses $H_0^{\{\ell+1\}}, \ldots, H_0^{\{d\}}$, according to the scheme.

Therefore, applying (11) to $\Lambda_{T_2}^{(j)}$, we obtain

$$\begin{aligned} FWER_{II} &= \boldsymbol{P}_{\mathcal{T}}\left\{\Lambda_{T_2}^{\{\ell\}} \le b_\ell\right\} \le \boldsymbol{P}_{\mathcal{T}}\left\{\Lambda_{T_2}^{\{\ell\}} \le b_{d-|\mathcal{F}|+1}\right\} \\ &= \boldsymbol{P}_{\mathcal{T}}\left\{\min_{j\in\mathcal{F}} \Lambda_{T_2}^{(j)} \le b_{d-|\mathcal{F}|+1}\right\} \\ &\le \sum_{j\in\mathcal{F}} \boldsymbol{P}_{\mathcal{T}}\left\{\Lambda_{T_2}^{(j)} \le \log(\beta/|\mathcal{F}|)\right\} \le \sum_{j\in\mathcal{F}} \beta/|\mathcal{F}| = \beta. \end{aligned}$$

5.5. Proof of Theorem 4

The intersection scheme satisfies all the conditions of Theorem 2, therefore, the stopping time T^* is proper, and the scheme controls $FWER_I \le \alpha$. At the same time, it satisfies Theorem 3, and therefore, it controls $FWER_{II} \le \beta$.

6. Acknowledgements

The authors are grateful to the Editor Professor N. Balakrishnan, the Associate Editor, and to the anonymous referee for deep, insightful, and encouraging comments that helped us tremendously. Research of both authors is funded by the National Science Foundation grant DMS 1007775. Research of the second author is partially supported by the National Security Agency grant H98230-11-1-0147. This funding is greatly appreciated.

References

[1] C. Jennison and B. W. Turnbull, Group sequential tests for bivariate response: Interim analyses of clinical trials with both efficacy and safety endpoints, Biometrics 49 (1993) 741–752.

[2] P. C. O'Brien, Procedures for comparing samples with multiple endpoints, Biometrics 40 (1984) 1079–1087.

[3] S. J. Pocock, N. L. Geller, and A. A. Tsiatis, The analysis of multiple endpoints in clinical trials, Biometrics 43 (1987) 487–498.

[4] D. H. Baillie, Multivariate acceptance sampling - some applications to defence procurement, The Statistician 36 (1987) 465–478.

[5] D. C. Hamilton, M. L. Lesperance, A consulting problem involving bivariate acceptance sampling by variables, Canadian J. Stat. 19 (1991) 109–117.

[6] A. G. Tartakovsky, V. V. Veeravalli, Change-point detection in multichannel and distributed systems with applications, in: N. Mukhopadhyay, S. Datta and S. Chattopadhyay, Eds., Applications of Sequential Methodologies, Marcel Dekker, Inc., New York, 2004, pp. 339–370.

[7] A. G. Tartakovsky, X. R. Li, and G. Yaralov, Sequential detection of targets in multichannel systems, IEEE Trans. Information Theory 49 (2) (2003) 425–445.

[8] S. Dudoit, J. P. Shaffer, and J. C. Boldrick, Multiple hypothesis testing in microarray experiment, Stat. Science 18 (2003) 71–103.

[9] J. P. Shaffer, Multiple hypothesis testing, Annual Review of Psychology 46 (1995) 561–584.

[10] M. Ghosh, N. Mukhopadhyay and P. K. Sen, Sequential Estimation, Wiley, New York, 1997.

[11] R. A. Betensky, An O'Brien-Fleming sequential trial for comparing three treatments, Ann. Stat. 24 (4) (1996) 1765–1791.

[12] D. Edwards, Extended-Paulson sequential selection, Ann. Stat. 15 (1) (1987) 449–455.

[13] D. G. Edwards, J. C. Hsu, Multiple comparisons with the best treatment, J. Amer. Stat. Assoc. 78 (1983) 965–971.

[14] M. D. Hughes, Stopping guidelines for clinical trials with multiple treatments, Statistics in Medicine 12 (1993) 901–915.

[15] C. Jennison and B. W. Turnbull, Group sequential methods with applications to clinical trials, Chapman & Hall, Boca Raton, FL, 2000.

[16] P. C. O'Brien, T. R. Fleming, A multiple testing procedure for clinical trials, Biometrika 35 (1979) 549–556.

[17] D. Siegmund, A sequential clinical trial for comparing three treatments, Ann. Stat. 21 (1993) 464–483.

[18] R. R. Wilcox, Extention of Hochberg's two-stage multiple comparison method, in: N. Mukhopadhyay, S. Datta and S. Chattopadhyay, Eds., Applications of Sequential Methodologies, Marcel Dekker, Inc., New York, 2004, pp. 371–380.

[19] S. Zacks, Stage-wise Adaptive Designs, Wiley, Hoboken, NJ, 2009.

[20] E. Paulson, A sequential procedure for comparing several experimental categories with a standard or control, Ann. Math. Stat. 33 (1962) 438–443.

[21] C. Jennison, I. M. Johnstone, and B. W. Turnbull, Asymptotically optimal procedures for sequential adaptive selection of the best of several normal means, in: S. S. Gupta and J. O. Berger, eds., Statistical Decision Theory and Related Topics III, Vol. 2, Academic Press, New York, 1982, pp. 55–86.

[22] E. Paulson, A sequential procedure for selecting the population with the largest mean from k normal populations, Ann. Math. Stat. 35 (1964) 174–180.

[23] P. Armitage, Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis, J. Roy. Statist. Soc. B 12 (1950) 137–144.

[24] C. W. Baum, V. V. Veeravalli, A sequential procedure for multihypothesis testing, IEEE Trans. Inform. Theory 40 (1994) 1994–2007.

[25] A. Novikov, Optimal sequential multiple hypothesis tests, Kybernetika 45 (2) (2009) 309–330.

[26] G. Simons, Lower bounds for the average sample number of sequential multihypothesis tests, Ann. Math. Stat. 38 (5) (1967) 1343–1364.

[27] Z. Govindarajulu, Sequential Statistics, World Scientific Publishing Co, Singapore, 2004.

[28] A. Wald, Sequential Analysis, Wiley, New York, 1947.

[29] A. Wald, J. Wolfowitz, Optimal character of the sequential probability ratio test, Ann. Math. Statist. 19 (1948) 326–339.

[30] D. Siegmund, Sequential Analysis: Tests and Confidence Intervals, Springer-Verlag, New York, 1985.

[31] M. Sobel, A. Wald, A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution, Ann. Math. Stat. 20 (4) (1949) 502–522.

[32] V. P. Dragalin, A. G. Tartakovsky, and V. V. Veeravalli, Multihypothesis sequential probability ratio tests. Part I: Asymptotic optimality, IEEE Trans. Inform. Theory 45 (7) (1999) 2448–2461.

[33] T. L. Lai, Sequential multiple hypothesis testing and efficient fault detection-isolation in stochastic systems, IEEE Trans. Inform. Theory 46 (2) (2000) 595–608.

[34] E. Glimm, W. Maurer, and F. Bretz, Hierarchical testing of multiple endpoints in group-sequential trials, Statistics in Medicine 29 (2010) 219–228.

[35] A. C. Tamhane, C. R. Mehta, and L. Liu, Testing a primary and a secondary endpoint in a group sequential design, Biometrics 66 (2010) 1174–1184.

[36] W. Maurer, E. Glimm, and F. Bretz, Multiple and repeated testing of primary, coprimary, and secondary hypotheses, Statistics in Biopharmaceutical Research 3 (2) (2011) 336–352.

[37] D.-I. Tang and N. L. Geller, Closed testing procedures for group sequential clinical trials with multiple endpoints, Biometrics 55 (1999) 1188–1192.

[38] J. Bartroff and T.-L. Lai, Multistage tests of multiple hypotheses, Communications in Statistics - Theory and Methods 39 (2010) 1597–1607.

[39] Y. Benjamini, F. Bretz, and S. Sarkar, Eds., Recent Developments in Multiple Comparison Procedures, IMS Lecture Notes - Monograph Series, Beachwood, Ohio, 2004.

[40] S. Dudoit, M. J. van der Laan, Multiple Testing Procedures with Applications to Genomics, Springer, New York, 2008.

[41] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. Royal Stat. Soc. 57 (1) (1995) 289–300.

[42] Y. Hochberg, A. C. Tamhane, Multiple comparison procedures, Wiley, New York, 1987.

[43] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Stat. 6 (1979) 65–70.

[44] S. K. Sarkar, Some probability inequalities for ordered mtp_2 random variables: a proof of the simes conjecture, Ann. Stat. 26 (2) (1998) 494–504.

[45] S. K. Sarkar, Some results on false discovery rate in stepwise multiple testing procedures, Ann. Stat. 30 (1) (2002) 239–257.

[46] Z. Sidak, Rectangular confidence regions for the means of multivariate normal distributions, J. Amer. Statist. Assoc. 62 (1967) 626–633.

[47] R. J. Simes, An improved Bonferroni procedure for multiple tests of significance, Biometrika 73 (1986) 751–754.

[48] E. L. Lehmann, J. P. Romano, Generalizations of the familywise error rate, Ann. Stat. 33 (2005) 1138–1154.

[49] J. P. Romano, M. Wolf, Control of generalized error rates in multiple testing, Ann. Stat. 35 (2007) 1378–1408.

[50] S. K. Sarkar, Step-up procedures controlling generalized FWER and generalized FDR, Ann. Stat. 35 (2007) 2405–2420.

[51] S. K. Sarkar, W. Guo, On a generalized false discovery rate, Ann. Stat. 37 (3) (2009) 1545–1565.

[52] G. Casella and R. L. Berger, Statistical Inference, Duxbury Press, Belmont, CA, 2002.

[53] Z. Govindarajulu, The Sequential Statistical Analysis of Hypothesis Testing, Point and Interval Estimation, and Decision Theory, American Sciences Press, Columbus, Ohio, 1987.

[54] M. Basseville, I. V. Nikiforov, Detection of Abrupt Changes: Theory and Application, PTR Prentice-Hall, Inc., 1993.

[55] S. De and M. Baron, Sequential bonferroni methods for multiple hypothesis testing with strong control of familywise error rates I and II, Sequential Analysis (in press).

[56] J. Neter, M. Kutner, C. Nachtsheim and W. Wasserman, Applied Linear Statistical Models, 4th ed., McGraw-Hill, 1996.

[57] D. Williams, Probability with Martingales, Cambridge University Press, Cambridge, UK, 1991.

[58] H. H. Kuo, Introduction to Stochastic Integration, Springer, New York, 2006.