The Annals of Statistics
2019, Vol. 47, No. 3, 1776–1803
https://doi.org/10.1214/18-AOS1737

SEQUENTIAL MULTIPLE TESTING WITH GENERALIZED ERROR CONTROL: AN ASYMPTOTIC OPTIMALITY THEORY[1]

BY YANGLEI SONG AND GEORGIOS FELLOURIS

University of Illinois, Urbana–Champaign

The sequential multiple testing problem is considered under two generalized error metrics. Under the first one, the probability of at least k mistakes, of any kind, is controlled. Under the second, the probabilities of at least k_1 false positives and at least k_2 false negatives are simultaneously controlled. For each formulation, the optimal expected sample size is characterized, to a first-order asymptotic approximation as the error probabilities go to 0, and a novel multiple testing procedure is proposed and shown to be asymptotically efficient under every signal configuration. These results are established when the data streams for the various hypotheses are independent and each local log-likelihood ratio statistic satisfies a certain strong law of large numbers. In the special case of i.i.d. observations in each stream, the gains of the proposed sequential procedures over fixed-sample size schemes are quantified.

1. Introduction. In the early development of multiple testing, the focus was on procedures that control the probability of at least *one* false positive, that is, falsely rejected null [13, 14, 22]. As this requirement can be prohibitive when the number of hypotheses is large, the emphasis gradually shifted to the control of less stringent error metrics, such as (i) the expectation [4] or the quantiles [18] of the *false discovery proportion*, that is, the proportion of false positives among the rejected nulls, and (ii) the *generalized familywise error rate*, that is, the probability of at least $k \geq 1$ false positives [15, 18]. During the last two decades, various procedures have been proposed to control the above error metrics [5, 12, 25, 26]. Further, the problem of maximizing the number of true positives subject to a generalized control on false positives has been studied in [19, 23, 30, 31], whereas in [6] the false negatives are incorporated into the risk function in a Bayesian decision theoretic framework.

In all previous references, it is assumed that the sample size is deterministic. However, in many applications data are collected in real time and a reliable decision needs to be made as quickly as possible. Such applications fall into the

Received August 2016; revised April 2018.

[1]Supported in part by NSF Grants CCF-1514245 and DMS-1737962 and in part by the Simons Foundation under Grant C3663.

MSC2010 subject classifications. 62L10.

Key words and phrases. Multiple testing, sequential analysis, asymptotic optimality, generalized familywise error rates, misclassification rate.

framework of *sequential hypothesis testing*, which was introduced in the groundbreaking work of Wald [35] and has been studied extensively since then (see, e.g., [32]).

When testing simultaneously *multiple* hypotheses with data collected from a different stream for each hypothesis, there are two natural generalizations of Wald's sequential framework. In the first one, sampling can be terminated earlier in some data streams [1, 3, 21]. In the second, which is the focus of this paper, sampling is terminated at the same time in all streams [7, 8]. The latter setup is motivated by applications such as multichannel signal detection [34], multiple access wireless network [24] and multisensor surveillance systems [11], where a centralized decision maker needs to make a decision regarding the presence or absence of signal, for example, an intruder, in multiple channels/areas monitored by a number of sensors. This framework is also motivated by online surveys and crowdsourcing tasks [17], where the goal is to find "correct" answers to a fixed number of questions, for example, regarding some product or service, by asking the smallest necessary number of people.

In this paper, we focus on two related, yet distinct, generalized error metrics. The first one is a generalization of the usual misclassification rate [20, 21], where the probability of at least $k \geq 1$ mistakes, of any kind, is controlled. The second one controls generalized familywise error rates of both types [1, 9], that is, the probabilities of at least $k_1 \geq 1$ false positives and at least $k_2 \geq 1$ false negatives.

Various sequential procedures have been proposed recently to control such generalized familywise error rates [1–3, 7–9]. To the best of our knowledge, the efficiency of these procedures is understood only in the case of *classical* familywise error rates, that is, when $k_1 = k_2 = 1$. Specifically, in the case of independent streams with i.i.d. observations, an asymptotic lower bound was obtained in [28] for the optimal expected sample size (ESS) as the error probabilities go to 0, and was shown to be attained, under any signal configuration, by several existing procedures. However, the results in [28] do not extend to *generalized* error metrics, since the technique for the proof of the asymptotic lower bound requires that the probability of not identifying the correct subset of signals goes to 0. Further, as we shall see, existing procedures fail to be asymptotically optimal, in general, under generalized error metrics.

The lack of an optimality theory under such generalized error control also implies that it is not well understood how the best possible ESS depends on the user-specified parameters. This limits the applicability of generalized error metrics, as it is not clear for the practitioner how to select the number of hypotheses to be "sacrificed" for the sake of a faster decision.

In this paper, we address this research gap by developing an asymptotic optimality theory for the sequential multiple testing problem under the two generalized error metrics mentioned above. Specifically, for each formulation we characterize the optimal ESS as the error probabilities go to 0, and propose a novel, feasible sequential multiple testing procedure that achieves the optimal ESS under

TABLE 1
Procedures marked with † are new. Procedures in bold font are asymptotically optimal (AO) without requiring a special testing structure. GMIS is short for generalized misclassification rate, and GFWER for generalized familywise error rates

Procedure	Metric	Section	Main results	Conditions for AO
Sum-Intersection†	GMIS	3.1	Theorem 3.3	(8)
Leap†	GFWER	4.2	Theorem 4.3	(8)
Asym. Sum-Intersection†	GFWER	4.1	Corollary 4.4	(8) + (11) + (12)
Intersection	Both	2.2	Corollary 3.4/4.4	(8) + (11)/(12)
MNP (fixed-sample)	Both	2.3	Theorem 3.5/4.5	Not optimal

every signal configuration. These results are established under the assumption of independent data streams, and require that the log-likelihood ratio statistic in each stream satisfies a certain strong law of large numbers. Thus, even in the case of classical familywise error rates, we extend the corresponding results in [28] by relaxing the i.i.d. assumption in each stream.

Finally, whenever sequential testing procedures are utilized, it is of interest to quantify the savings in the ESS over fixed-sample size schemes with the same error control guarantees. In the case of i.i.d. data streams, we obtain an asymptotic lower bound for the gains of sequential sampling over *any* fixed-sample size scheme, and also characterize the asymptotic gains over a specific fixed-sample size procedure.

In order to convey the main ideas and results with the maximum clarity, we first consider the case that the local hypotheses are simple, and then extend our results to the case of composite hypotheses. Thus, the remainder of the paper is organized as follows: in Section 2, we formulate the two problems of interest in the case of simple hypotheses. The case of generalized misclassification rate is presented in Section 3, and the case of generalized familywise error rates in Section 4. In Section 5, we present two simulation studies under the second error metric. In Section 6, we extend our results to the case of composite hypotheses. We conclude and discuss potential extensions of this work in Section 7. Proofs are presented in the Appendix (Supplementary Material [29]), where we also present more simulation studies and a detailed analysis of the case of composite hypotheses. For convenience, we list in Table 1 the procedures that are considered in this work.

2. Problem formulation. Consider *independent* streams of observations, $X^j := \{X^j(n) : n \in \mathbb{N}\}$, where $j \in [J] := \{1, \dots, J\}$ and $\mathbb{N} := \{1, 2, \dots\}$. For each $j \in [J]$, we denote by P^j the distribution of X^j and consider two simple hypotheses for it,

$$\mathsf{H}_0^j : \mathsf{P}^j = \mathsf{P}_0^j \quad \text{versus} \quad \mathsf{H}_1^j : \mathsf{P}^j = \mathsf{P}_1^j. \tag{1}$$

We denote by P_A the distribution of $(X^1, \dots, X^J)$ when $A \subset [J]$ is the subset of data streams with signal, that is, in which the alternative hypothesis is correct. Due

to the assumption of independence among streams, P_A is the following product measure:

$$\mathsf{P}_A := \bigotimes_{j=1}^{J} \mathsf{P}^j; \qquad \mathsf{P}^j = \begin{cases} \mathsf{P}_0^j & \text{if } j \notin A, \\ \mathsf{P}_1^j & \text{if } j \in A. \end{cases} \tag{2}$$

Moreover, we denote by $\mathcal{F}_n^j$ the σ-field generated by the first n observations in the jth stream, that is, $\sigma(X^j(1), \ldots, X^j(n))$, and by $\mathcal{F}_n$ the σ-field generated by the first n observations in all streams, that is, $\sigma(\mathcal{F}_n^j,\ j \in [J])$, where $n \in \mathbb{N}$.

Assuming that the data in all streams become available *sequentially*, the goal is to stop sampling *as soon as possible*, and upon stopping to solve the J hypothesis testing problems subject to certain error control guarantees. Formally, a *sequential multiple testing procedure* is a pair $\delta = (T, D)$ where T is an $\{\mathcal{F}_n\}$-stopping time at which sampling is terminated in all streams, and D an $\mathcal{F}_T$-measurable, J-dimensional vector of Bernoullis, $(D^1, \ldots, D^J)$, so that the alternative hypothesis is selected in the jth stream if and only if $D^j = 1$. With an abuse of notation, we also identify D with the subset of streams in which the alternative hypothesis is selected upon stopping, that is, $\{j \in [J] : D^j = 1\}$.

We consider two kinds of error control, which lead to two different problems. Their main difference is that the first one does not differentiate between *false positives*, that is, rejecting the null when it is correct, and *false negatives*, that is, accepting the null when it is false. Specifically, in the first one we control the generalized misclassification rate, that is, the probability of committing *at least k mistakes, of any kind*, where k is a user-specified integer such that $1 \le k < J$. When A is the true subset of signals, a decision rule D makes at least k mistakes, of any kind, if D and A differ in at least k components, that is, $|A \triangle D| \ge k$, where for any two sets A and D, $A \triangle D$ is their symmetric difference, that is, $(A \setminus D) \cup (D \setminus A)$, and $|\cdot|$ denotes set-cardinality. Thus, given tolerance level $\alpha \in (0, 1)$, the class of multiple testing procedures of interest in this case is

$$\Delta_k(\alpha) := \Big\{(T, D) : \max_{A \subset [J]} \mathsf{P}_A\big(|A \triangle D| \ge k\big) \le \alpha\Big\}.$$

Then the first problem is formulated as follows.

PROBLEM 2.1. Given a user-specified integer k in $[1, J)$, find a sequential multiple testing procedure that (i) controls the generalized misclassification rate, that is, it can be designed to belong to $\Delta_k(\alpha)$ for any given α, and (ii) achieves the smallest possible expected sample size,

$$N_A^*(k, \alpha) := \inf_{(T,D) \in \Delta_k(\alpha)} \mathsf{E}_A[T]$$

for every $A \subset [J]$, to a first-order asymptotic approximation as $\alpha \to 0$.

In the second problem of interest in this work, we control generalized familywise error rates of both types, that is, the probabilities of *at least k_1 false positives and at least k_2 false negatives*, where $k_1, k_2 \geq 1$ are integers such that $k_1 + k_2 \leq J$. When the true subset of signals is A, a decision rule D makes at least k_1 false positives when $|D \setminus A| \geq k_1$ and at least k_2 false negatives when $|A \setminus D| \geq k_2$. Thus, given tolerance levels $\alpha, \beta \in (0, 1)$, the class of procedures of interest in this case is

$$\begin{aligned}\Delta_{k_1,k_2}(\alpha,\beta) := \Big\{(T,D) : \max_{A\subset[J]} \mathsf{P}_A\big(|D\setminus A| \geq k_1\big) \leq \alpha \text{ and } \\ \max_{A\subset[J]} \mathsf{P}_A\big(|A\setminus D| \geq k_2\big) \leq \beta\Big\}.\end{aligned} \tag{3}$$

Then the second problem is formulated as follows.

PROBLEM 2.2. Given user-specified integers $k_1, k_2 \geq 1$ such that $k_1 + k_2 \leq J$, find a sequential multiple testing procedure that (i) controls generalized familywise error rates of both types, that is, it can be designed to belong to $\Delta_{k_1,k_2}(\alpha, \beta)$ for any given $\alpha, \beta \in (0, 1)$, and (ii) achieves the smallest possible expected sample size,

$$N_A^*(k_1, k_2, \alpha, \beta) := \inf_{(T,D)\in\Delta_{k_1,k_2}(\alpha,\beta)} \mathsf{E}_A[T]$$

for every $A \subset [J]$, to a first-order asymptotic approximation as α and β go to 0, at arbitrary rates.

2.1. *Assumptions.* We now state the assumptions that we will make in the next two sections in order to solve these two problems. First of all, for each $j \in [J]$ we assume that the probability measures P_0^j and P_1^j in (1) are mutually absolutely continuous when restricted to $\mathcal{F}_n^j$, and we denote the corresponding log-likelihood ratio (LLR) statistic as follows:

$$\lambda^j(n) := \log \frac{d\mathsf{P}_1^j}{d\mathsf{P}_0^j}(\mathcal{F}_n^j) \qquad \text{for } n \in \mathbb{N}.$$

For $A, C \subset [J]$ and $n \in \mathbb{N}$, we denote by $\lambda^{A,C}(n)$ the LLR of P_A versus P_C when both measures are restricted to $\mathcal{F}_n$, and from (2) it follows that

$$\lambda^{A,C}(n) := \log \frac{d\mathsf{P}_A}{d\mathsf{P}_C}(\mathcal{F}_n) = \sum_{j\in A\setminus C} \lambda^j(n) - \sum_{j\in C\setminus A} \lambda^j(n). \tag{4}$$

In order to guarantee that the proposed multiple testing procedures terminate almost surely and satisfy the desired error control, it will suffice to assume that

$$\mathsf{P}_1^j\Big(\lim_{n\to\infty} \lambda^j(n) = \infty\Big) = \mathsf{P}_0^j\Big(\lim_{n\to\infty} \lambda^j(n) = -\infty\Big) = 1 \qquad \forall j \in [J]. \tag{5}$$

In order to establish an asymptotic lower bound on the optimal ESS for each problem, we will need the stronger assumption that for each $j \in [J]$ there are positive numbers, $\mathcal{I}_1^j, \mathcal{I}_0^j$, such that the following Strong Laws of Large Numbers (SLLN) hold:

$$\mathsf{P}_1^j\left(\lim_{n\to\infty} \frac{\lambda^j(n)}{n} = \mathcal{I}_1^j\right) = \mathsf{P}_0^j\left(\lim_{n\to\infty} \frac{\lambda^j(n)}{n} = -\mathcal{I}_0^j\right) = 1. \tag{6}$$

When the LLR statistic in each stream has *independent and identically distributed (i.i.d.) increments*, the SLLN (6) will also be sufficient for establishing the asymptotic optimality of the proposed procedures. When this is not the case, we will need an assumption on the rate of convergence in (6). Specifically, we will assume that for every $\varepsilon > 0$ and $j \in [J]$,

$$\sum_{n=1}^{\infty} \mathsf{P}_1^j\left(\left|\frac{\lambda^j(n)}{n} - \mathcal{I}_1^j\right| > \varepsilon\right) < \infty, \qquad \sum_{n=1}^{\infty} \mathsf{P}_0^j\left(\left|\frac{\lambda^j(n)}{n} + \mathcal{I}_0^j\right| > \varepsilon\right) < \infty. \tag{7}$$

Condition (7) is known as *complete* convergence [16], and is a stronger assumption than (6), due to the Borel–Cantelli lemma. This condition is satisfied in various testing problems where the observations in each data stream are dependent, such as autoregressive time-series models and state-space models. For more details, we refer to [32], Chapter 3.4.

To sum up, the only distributional assumption for our asymptotic optimality theory is that the LLR statistic in each stream:

$$\begin{aligned} &\text{either has i.i.d. increments and satisfies the SLLN (6),}\\ &\text{or satisfies the SLLN with complete convergence (7).} \end{aligned} \tag{8}$$

REMARK 2.1. If (6) (resp., (7)) holds, the normalized LLR, $\lambda^{A,C}(n)/n$, defined in (4), converges almost surely (resp. completely) under P_A to

$$\mathcal{I}^{A,C} := \sum_{i\in A\setminus C} \mathcal{I}_1^i + \sum_{j\in C\setminus A} \mathcal{I}_0^j. \tag{9}$$

The numbers $\mathcal{I}^{A,C}$ and $\mathcal{I}^{C,A}$ will turn out to determine the inherent difficulty in distinguishing between P_A and P_C and will play an important role in characterizing the optimal performance under P_A and P_C, respectively.

2.2. *The Intersection rule.* To the best of our knowledge, Problem 2.2 has been solved only under the assumption of i.i.d. data streams and *only in the case of classical error control, that is, when* $k_1 = k_2 = 1$ [28]. An asymptotically optimal procedure in this setup is the so-called "*Intersection*" rule, $\delta_I := (T_I, D_I)$, proposed in [7, 8], where

$$\begin{aligned} T_I &:= \inf\{n \geq 1 : \lambda^j(n) \notin (-a, b) \text{ for every } j \in [J]\},\\ D_I &:= \{j \in [J] : \lambda^j(T_I) > 0\}, \end{aligned} \tag{10}$$

and a, b are positive thresholds. This procedure requires the local test statistic in *every* stream to provide sufficiently strong evidence for the sampling to be terminated. The Intersection rule was also shown in [9] to control *generalized* family-wise error rates, however its efficiency in this setup remains an open problem, even in the case of i.i.d. data streams. Our asymptotic optimality theory in the next sections will reveal that the Intersection rule is asymptotically optimal with respect to Problems 2.1 and 2.2 only when the multiple testing problem satisfies *a very special structure*.

DEFINITION 2.1. We say that the multiple testing problem (1) is:

(i) *symmetric*, if for every $j \in [J]$ the distribution of λ^j under P_0^j is the same as the distribution of $-\lambda^j$ under P_1^j,

(ii) *homogeneous*, if for every $j \in [J]$ the distribution of λ^j under P_i^j does not depend on j, where $i \in \{0, 1\}$.

It is clear that when the multiple testing problem is both *symmetric and homogeneous*, we have

$$\mathcal{I}_0^j = \mathcal{I}_1^j = \mathcal{I} \qquad \text{for every } j \in [J]. \tag{11}$$

In the next sections, we will show that the Intersection rule is asymptotically optimal for Problem 2.1 when (11) holds, whereas its asymptotic optimality with respect to Problem 2.2 will *additionally* require that the user-specified parameters satisfy the following conditions:

$$k_1 = k_2 \quad \text{and} \quad \alpha = \beta. \tag{12}$$

2.3. *Fixed-sample size schemes*. Let $\Delta_{\text{fix}}(n)$ denote the class of procedures for which the decision rule depends on the data collected up to a *deterministic* time n, that is,

$$\Delta_{\text{fix}}(n) := \big\{(n, D) : D \subset [J] \text{ is } \mathcal{F}_n\text{-measurable}\big\}.$$

For any given integers $k, k_1, k_2 \geq 1$ with $k, k_1 + k_2 < J$ and $\alpha, \beta \in (0, 1)$, let

$$\begin{aligned} n^*(k, \alpha) &:= \inf\big\{n \in \mathbb{N} : \Delta_{\text{fix}}(n) \cap \Delta_k(\alpha) \neq \varnothing\big\}, \\ n^*(k_1, k_2, \alpha, \beta) &:= \inf\big\{n \in \mathbb{N} : \Delta_{\text{fix}}(n) \cap \Delta_{k_1,k_2}(\alpha, \beta) \neq \varnothing\big\}, \end{aligned} \tag{13}$$

denote the minimum sample sizes required by *any fixed-sample size scheme* under the two error metrics of interest. In the case of i.i.d. observations in the data streams, we establish *asymptotic lower bounds* for the above two quantities as the error probabilities go to 0. To the best of our knowledge, there is no fixed-sample size procedure that attains these bounds. For this reason, we also study a specific

procedure that runs a *Neyman–Pearson test at each stream*. Formally, this procedure is defined as follows:

$$\delta_{\mathrm{NP}}(n,h) := \big(n, D_{\mathrm{NP}}(n,h)\big), \qquad D_{\mathrm{NP}}(n,h) := \big\{j \in [J] : \lambda^j(n) > nh_j\big\}, \tag{14}$$

where $h = (h_1, \ldots, h_J) \in \mathbb{R}^J$, $n \in \mathbb{N}$, and we refer to it as *multiple Neyman–Pearson (MNP) rule*. In the case of Problem 2.1, we characterize the minimum sample size required by this procedure,

$$n_{\mathrm{NP}}(k,\alpha) := \inf\big\{n \in \mathbb{N} : \exists h \in \mathbb{R}^J, \delta_{\mathrm{NP}}(n,h) \in \Delta_k(\alpha)\big\},$$

to a first-order approximation as $\alpha \to 0$. In the case of Problem 2.2, for simplicity of presentation we further restrict ourselves to *homogeneous*, but not necessarily symmetric, multiple testing problems, and characterize the asymptotic minimum sample size required by the MNP rule that utilizes the same threshold in each stream, that is,

$$\hat{n}_{\mathrm{NP}}(k_1,k_2,\alpha,\beta) := \inf\big\{n \in \mathbb{N} : \exists h \in \mathbb{R}, \delta_{\mathrm{NP}}(n, h\mathbf{1}_J) \in \Delta_{k_1,k_2}(\alpha,\beta)\big\},$$

where $\mathbf{1}_J \in \mathbb{R}^J$ is a J-dimensional vector of ones.

2.4. *The i.i.d. case*. As mentioned earlier, our asymptotic optimality theory will apply whenever condition (8) holds, thus, beyond the case of i.i.d. data streams. However, our analysis of *fixed-sample size* schemes will rely on large deviation theory [10] and will be focused on the i.i.d. case. Thus, it is useful to introduce some relevant notation for this setup.

Specifically, when for each $j \in [J]$ the observations in the jth stream are independent with common density f^j relative to a σ-finite measure ν^j, the hypothesis testing problem (1) takes the form

$$\mathsf{H}_0^j : f^j = f_0^j \quad \text{versus} \quad \mathsf{H}_1^j : f^j = f_1^j, \tag{15}$$

and $\mathcal{I}_1^j, \mathcal{I}_0^j$ correspond to the *Kullback–Leibler divergences* between f_1^j and f_0^j, that is,

$$\mathcal{I}_1^j = \int \log(f_1^j/f_0^j) f_1^j \, d\nu^j, \qquad \mathcal{I}_0^j = \int \log(f_0^j/f_1^j) f_0^j \, d\nu^j. \tag{16}$$

In this case, each LLR statistic λ^j has i.i.d. increments, and (8) is satisfied as long as $\mathcal{I}_1^j$ and $\mathcal{I}_0^j$ are both positive and finite. For each $j \in [J]$, we further introduce the convex conjugate of the cumulant generating function of $\lambda^j(1)$

$$z \in \mathbb{R} \mapsto \Phi^j(z) := \sup_{\theta \in \mathbb{R}} \{z\theta - \Psi^j(\theta)\} \qquad \text{where } \Psi^j(\theta) := \log \mathsf{E}_0^j[e^{\theta\lambda^j(1)}]. \tag{17}$$

The value of Φ^j at zero is the *Chernoff information* [10] for the testing problem (15), and we will denote it as $\mathcal{C}^j$, that is, $\mathcal{C}^j := \Phi^j(0)$.

Finally, we will illustrate our general results in the case of testing normal means. Hereafter, $\mathcal{N}$ denotes the density of the normal distribution.

EXAMPLE 2.1. If $f_0^j = \mathcal{N}(0, \sigma_j^2)$ and $f_1^j = \mathcal{N}(\mu_j, \sigma_j^2)$ for all $j \in [J]$, then

$$\lambda^j(1) = \theta_j^2\big(X^j(1)/\mu_j - 1/2\big) \qquad \text{where } \theta_j := \mu_j/\sigma_j.$$

Consequently, the multiple testing problem is symmetric and

$$\mathcal{I}^j := \mathcal{I}_0^j = \mathcal{I}_1^j = \theta_j^2/2, \qquad \Phi^j(z) = \big(z + \mathcal{I}^j\big)^2/\big(4\mathcal{I}^j\big) \qquad \text{for any } z \in \mathbb{R}. \tag{18}$$

2.5. *Notation*. We collect here some notation that will be used extensively throughout the rest of the paper: C_k^J denotes the binomial coefficient $\binom{J}{k}$, that is, the number of subsets of size k from a set of size J; $a \vee b$ represents $\max\{a, b\}$; $x \sim y$ means that $\lim_y x/y = 1$ and $x(b) = o(1)$ that $\lim_b x(b) = 0$, with $y, b \to 0$ or ∞. Moreover, we recall that $|\cdot|$ denotes set-cardinality, $\mathbb{N} := \{1, 2, \ldots\}$, $[J] := \{1, \ldots, J\}$, and that $A \triangle B$ is the symmetric difference, $(A \setminus B) \cup (B \setminus A)$, of two sets A and B.

3. Generalized misclassification rate. In this section, we consider Problem 2.1 and carry out the following program: first, we propose a novel procedure that controls the generalized misclassification rate. Then we establish an asymptotic lower bound on the optimal ESS and show that it is attained by the proposed scheme. As a corollary, we show that the Intersection rule is asymptotically optimal when condition (11) holds. Finally, we make a comparison with fixed-sample size procedures in the i.i.d. case (15).

3.1. *Sum-Intersection rule*. In order to implement the proposed procedure, which we will denote $\delta_S(b) := (T_S(b), D_S(b))$, we need at each time $n \in \mathbb{N}$ prior to stopping to order the *absolute values* of the local test statistics, $|\lambda^j(n)|$, $j \in [J]$. If we denote the corresponding ordered values by

$$\widetilde{\lambda}^1(n) \le \cdots \le \widetilde{\lambda}^J(n),$$

we can think of $\widetilde{\lambda}^1(n)$ (resp., $\widetilde{\lambda}^J(n)$) as the least (resp., most) "significant" local test statistic at time n, in the sense that it provides the weakest (resp., strongest) evidence in favor of either the null or the alternative. Then sampling is terminated at the first time the *sum of the k least significant local LLRs* exceeds some positive threshold b, and the null hypothesis is rejected in every stream that has a positive LLR upon stopping, that is,

$$T_S(b) := \inf\left\{n \ge 1 : \sum_{j=1}^{k} \widetilde{\lambda}^j(n) \ge b\right\}, \qquad D_S(b) := \big\{j \in [J] : \lambda^j\big(T_S(b)\big) > 0\big\}.$$

The threshold b is selected to guarantee the desired error control. When $k = 1$, $\delta_S(b)$ coincides with the Intersection rule, $\delta_I(b, b)$, defined in (10). When $k > 1$, the two rules are different but share a similar flavor, since $\delta_S(b)$ stops the first time n that *all sums* $\sum_{j \in B} |\lambda^j(n)|$ *with* $B \subset [J]$ *and* $|B| = k$ are simultaneously above b. For this reason, we refer to $\delta_S(b)$ as *Sum-Intersection rule*. Hereafter, we typically suppress the dependence of $\delta_S(b)$ on threshold b in order to lighten the notation.

3.2. *Error control of the Sum-Intersection rule.* For any choice of threshold b, the Sum-Intersection rule clearly terminates almost surely, under every signal configuration, as long as condition (5) holds. In the next theorem, we show how to select b to guarantee the desired error control. We stress that no additional distributional assumptions are needed for this purpose.

THEOREM 3.1. *Assume* (5) *holds. For any* $\alpha \in (0,1)$, *we have* $\delta_S(b_\alpha) \in \Delta_k(\alpha)$ *when*

$$b_\alpha = \bigl|\log(\alpha)\bigr| + \log\bigl(C_k^J\bigr). \tag{19}$$

PROOF. The proof can be found in Appendix B.1. □

The choice of b suggested by the previous theorem will be sufficient for establishing the asymptotic optimality of the Sum-Intersection rule, but may be conservative for practical purposes. In the absence of more accurate approximations for the error probabilities, we recommend finding the value of b for which the target level is attained using Monte Carlo simulation. This means simulating off-line, that is, before the sampling process begins, for every $A \subset [J]$ the error probability $\mathsf{P}_A(|A \triangle D_S(b)| \geq k)$ for various values of b, and then selecting the value for which the maximum of these probabilities over $A \subset [J]$ matches the nominal level α.

This simulation task is significantly facilitated when the multiple testing problem has a special structure. If the problem is *symmetric*, for any given threshold b the error probabilities of the Sum-Intersection rule coincide for all $A \subset [J]$; thus, it suffices to simulate the error probability under a single measure, for example, $\mathsf{P}_\varnothing$. If the problem is *homogeneous*, the error probabilities depend only on the size of A, not the actual subset; thus, it suffices to simulate the above probabilities for at most $J+1$ configurations. Similar ideas apply in the presence of blockwise homogeneity.

Moreover, it is worth pointing out that when b is large, importance sampling techniques can be applied to simulate the corresponding "small" error probabilities, similar to [27].

3.3. *Asymptotic lower bound on the optimal performance.* We now obtain an asymptotic (as $\alpha \to 0$) lower bound on $N_A^*(k,\alpha)$, the optimal ESS for Problem 2.1 when the true subset of signals is A, for any given $k \geq 1$. When $k = 1$, from [33], Theorem 2.2, it follows that when (6) holds, such a lower bound is given by $|\log(\alpha)| / \min_{C \neq A} \mathcal{I}^{A,C}$, where $\mathcal{I}^{A,C}$ is defined in (9). Thus, the asymptotic lower bound when $k = 1$ is determined by the "wrong" subset that is the most difficult to be distinguished from A, where the difficulty level is quantified by the information numbers defined in (9).

The techniques in [33] require that the probability of selecting the wrong subset goes to 0; thus, they do not apply to the case of generalized error control ($k > 1$).

Nevertheless, it is reasonable to conjecture that the corresponding asymptotic lower bound when $k > 1$ will still be determined by the wrong subset that is the most difficult to be distinguished from A, with the difference that a subset will now be "wrong" under P_A *if it differs from* A *in at least* k *components*, that is, if it does *not* belong to

$$\mathcal{U}_k(A) := \{C \subset [J] : |A \triangle C| < k\}.$$

This conjecture is verified by the following theorem.

THEOREM 3.2. *Fix* $k \geq 1$. *If* (6) *holds, then for any* $A \subset [J]$, *as* $\alpha \to 0$,

$$N_A^*(k,\alpha) \geq \frac{|\log(\alpha)|}{\mathcal{D}_A(k)}\big(1 - o(1)\big) \qquad \text{where } \mathcal{D}_A(k) := \min_{C \notin \mathcal{U}_k(A)} \mathcal{I}^{A,C}. \tag{20}$$

The proof in the case of the *classical* misclassification rate ($k = 1$) is based on a change of measure from P_A to P_{A^*}, where A^* is chosen such that (i) A is a "wrong" subset under P_{A^*}, that is, $A \neq A^*$ and (ii) A^* is "close" to A, in the sense that $\mathcal{I}^{A,A^*} \leq \mathcal{I}^{A,C}$ for every $C \neq A$ (see, e.g., [33], Theorem 2.2).

When $k \geq 2$, there are more than one "correct" subsets under P_A. The key idea in our proof is that for *each* "correct" subset $B \in \mathcal{U}_k(A)$ we apply a different change of measure $\mathsf{P}_A \to \mathsf{P}_{B^*}$, where B^* is chosen such that (i) B is a "wrong" subset under P_{B^*}, that is, $B \notin \mathcal{U}_k(B^*)$, and (ii) B^* is "close" to A, in the sense that $I^{A,B^*} \leq \mathcal{I}^{A,C}$ for every $C \notin \mathcal{U}_k(A)$. The existence of such B^* is established in Appendix B.2, and the proof of Theorem 3.2 is carried out in Appendix B.3.

3.4. *Asymptotic optimality.* We are now ready to establish the asymptotic optimality of the Sum-Intersection rule by showing that it attains the asymptotic lower bound of Theorem 3.2 under every signal configuration.

THEOREM 3.3. *Assume* (8) *holds. Then, for any* $A \subset [J]$ *we have as* $b \to \infty$ *that*

$$\mathsf{E}_A\big[T_S(b)\big] \leq \frac{b}{\mathcal{D}_A(k)}\big(1 + o(1)\big). \tag{21}$$

When in particular b *is selected such that* $\delta_S \in \Delta_k(\alpha)$ *and* $b \sim |\log(\alpha)|$, *for example, as in* (19), *then for every* $A \subset [J]$ *we have as* $\alpha \to 0$,

$$\mathsf{E}_A[T_S] \sim \frac{|\log \alpha|}{\mathcal{D}_A(k)} \sim N_A^*(k,\alpha).$$

PROOF. If (21) holds and b is such that $\delta_S \in \Delta_k(\alpha)$ and $b \sim |\log(\alpha)|$, then δ_S attains the asymptotic lower bound in Theorem 3.2. Thus, it suffices to prove (21), which is done in the Appendix B.4. □

The asymptotic characterization of the optimal ESS, $N_A^*(k,\alpha)$, illustrates the trade-off among the ESS, the number of mistakes to be tolerated, and the error tolerance level α. Specifically, it suggests that, for "small" values of α, tolerating $k-1$ mistakes reduces the ESS by a factor of $\mathcal{D}_A(k)/\mathcal{D}_A(1)$, which is *at least* k for every $A \subset [J]$. To justify the latter claim, note that if we denote the ordered information numbers $\{\mathcal{I}_1^j,\ j \in A\} \cup \{\mathcal{I}_0^j,\ j \notin A\}$ by $\widetilde{\mathcal{I}}^{(1)}(A) \le \cdots \le \widetilde{\mathcal{I}}^{(J)}(A)$, then

$$\mathcal{D}_A(k) = \sum_{j=1}^{k} \widetilde{\mathcal{I}}^{(j)}(A).$$

In the following corollary, we show that the Intersection rule is asymptotically optimal when (11) holds, which is the case for example when the multiple testing problem is *both symmetric and homogeneous*.

COROLLARY 3.4. (i) *Assume* (5) *holds. For any* $\alpha \in (0,1)$, *we have* $\delta_I(b,b) \in \Delta_k(\alpha)$ *when* b *is equal to* b_α/k, *where* b_α *is defined in* (19).

(ii) *Suppose* b *is selected such that* $\delta_I(b,b) \in \Delta_k(\alpha)$ *and* $b \sim |\log\alpha|/k$, *for example, as in* (*i*). *If* (8) *holds, then*

$$\mathsf{E}_A[T_I] \le \frac{|\log\alpha|}{k\mathcal{D}_A(1)}\big(1+o(1)\big).$$

If also (11) *holds, then for any* $A \subset [J]$ *we have as* $\alpha \to 0$ *that*

$$\mathsf{E}_A[T_I] \sim \frac{|\log\alpha|}{k\mathcal{I}} \sim N_A^*(k,\alpha).$$

PROOF. The proof can be found in Appendix B.5. □

REMARK 3.1. When (11) is violated, the Intersection rule fails to be asymptotically optimal. This will be illustrated with a simulation study in Appendix A.2.

3.5. *Fixed-sample size rules.* Finally, we focus on the i.i.d. case (15) and consider procedures that stop at a deterministic time, selected to control the generalized misclassification rate. We recall that $\mathcal{C}^j$ is the Chernoff information in the jth testing problem, and we denote by $\mathcal{B}(k)$ the sum of the smallest k local Chernoff informations, that is,

$$\mathcal{B}(k) := \sum_{j=1}^{k} \mathcal{C}^{(j)},$$

where $\mathcal{C}^{(1)} \le \mathcal{C}^{(2)} \le \cdots \le \mathcal{C}^{(J)}$ are the ordered values of the local Chernoff information numbers $\mathcal{C}^j$, $j \in [J]$.

THEOREM 3.5. *Consider the multiple testing problem with i.i.d. streams defined in* (15) *and suppose that the Kullback–Leibler numbers in* (16) *are positive and finite. For any user-specified integer* $1 \le k \le (J+1)/2$ *and* $A \subset [J]$, *we have as* $\alpha \to 0$

$$\frac{\mathcal{D}_A(k)}{\mathcal{B}(2k-1)}\big(1-o(1)\big) \le \frac{n^*(k,\alpha)}{N_A^*(k,\alpha)} \le \frac{n_{\mathrm{NP}}(k,\alpha)}{N_A^*(k,\alpha)} \sim \frac{\mathcal{D}_A(k)}{\mathcal{B}(k)}.$$

PROOF. The proof can be found in Appendix B.6. □

REMARK 3.2. Since any fixed time is also a stopping time, the lower bound is relevant only when $\mathcal{D}_A(k) > \mathcal{B}(2k-1)$ for some $A \subset [J]$.

We now specialize the results of the previous theorem to the *testing of normal means*, introduced in Example 2.1 (a Bernoulli example is presented in Appendix B.7). In this case, $\mathcal{C}^j = \mathcal{I}^j/4$ for every $j \in [J]$, which implies $\mathcal{D}_A(k) = 4\mathcal{B}(k)$ for every $A \subset [J]$, and by Theorem 3.5 it follows that

$$n_{\mathrm{NP}}(k,\alpha) \sim 4N_A^*(k,\alpha) \qquad \forall A \subset [J].$$

That is, for any $k \in [1,(J+1)/2]$, when utilizing the MNP rule instead of the proposed asymptotically optimal Sum-Intersection rule, the ESS increases by roughly a factor of 4, for small values of α, under every configuration. From Theorem 3.5, it also follows that for any $A \subset [J]$ we have

$$\liminf_{\alpha\to 0} \frac{n^*(k,\alpha)}{N_A^*(k,\alpha)} \ge \frac{4\mathcal{B}(k)}{\mathcal{B}(2k-1)}.$$

If in addition, the hypotheses have identical information numbers, that is, (11) holds, this lower bound is always larger than 2, which means that *any* fixed-sample size scheme will require at least twice as many observations as the Sum-Intersection rule, for small error probabilities.

4. Generalized familywise error rates *of both kinds*. In this section, we study Problem 2.2. While we follow similar ideas and the results are of similar nature as in the previous section, the proposed procedure and the proof of its asymptotic optimality turn out to be much more complicated.

To describe the proposed multiple testing procedure, we first need to introduce some additional notation. Specifically, we denote by

$$0 < \widehat{\lambda}^1(n) \le \cdots \le \widehat{\lambda}^{p(n)}(n)$$

the order statistics of the *positive* LLRs at time n, $\{\lambda^j(n) : \lambda^j(n) > 0, j \in [J]\}$, where $p(n)$ is the number of the strictly positive LLRs at time n. Similarly, we denote by

$$0 \le \check{\lambda}^1(n) \le \cdots \le \check{\lambda}^{q(n)}(n)$$

the order statistics of the absolute values of the *nonpositive* LLRs at time n, that is, $\{-\lambda^j(n) : \lambda^j(n) \le 0, j \in [J]\}$, where $q(n) := J - p(n)$. We also adopt the following convention:

$$\widehat{\lambda}^j(n) = \infty \qquad \text{if } j > p(n) \quad \text{and} \quad \check{\lambda}^j(n) = \infty \qquad \text{if } j > q(n). \tag{22}$$

Moreover, we use the following notation:

$$\lambda^{\widehat{i}_j(n)}(n) := \widehat{\lambda}^j(n) \qquad \forall j \in \{1, \dots, p(n)\},$$

$$\lambda^{\check{i}_j(n)}(n) := -\check{\lambda}^j(n) \qquad \forall j \in \{1, \dots, q(n)\}$$

for the indices of streams with *positive* and *nonpositive* LLRs at time n, respectively. Thus, stream $\widehat{i}_1(n)$ (resp., $\check{i}_1(n)$) has the least significant positive (resp., negative) LLR at time n.

4.1. *Asymmetric Sum-Intersection rule*. We start with a procedure that has the same decision rule as the Sum-Intersection procedure (Section 3.1), but a different stopping rule that accounts for the asymmetry in the error metric that we consider in this section. Specifically, we consider a procedure $\delta_0(a, b) \equiv (\tau_0, D_0)$ that stops as soon as the following two conditions are satisfied simultaneously: (i) the sum of the k_1 least significant positive LLRs is larger than $b > 0$, and (ii) the sum of the k_2 least significant negative LLRs is smaller than $-a < 0$. Formally,

$$\begin{aligned} \tau_0 &:= \inf\left\{ n \ge 1 : \sum_{j=1}^{k_1} \widehat{\lambda}^j(n) \ge b \text{ and } \sum_{j=1}^{k_2} \check{\lambda}^j(n) \ge a \right\}, \\ D_0 &:= \{ j \in [J] : \lambda^j(\tau_0) > 0 \} = \{ \widehat{i}_1(\tau_0), \dots, \widehat{i}_{p(\tau_0)}(\tau_0) \}. \end{aligned} \tag{23}$$

We refer to this procedure as *asymmetric Sum-Intersection rule*. Note that similarly to the Sum-Intersection rule, this procedure does not require strong evidence from every individual stream in order to terminate sampling. Indeed, upon stopping there may be insufficient evidence for the hypotheses that correspond to the $k_1 - 1$ least significant positive statistics and the $k_2 - 1$ least significant negative statistics, turning them into the anticipated false positives and false negatives, respectively, which we are allowed to make.

We will see that while the asymmetric Sum-Intersection rule can control generalized familywise error rates of both types, it is not in general asymptotically optimal. To understand why this is the case, let A denote true subset of streams with signals and suppose that there is a subset B of ℓ streams with *noise*, that is, $B \subset A^c$ with $|B| = \ell$, such that $\ell < k_1$ and

$$\mathcal{I}_1^j \gg \mathcal{I}_0^{i_1} \gg \mathcal{I}_0^{i_2} \qquad \forall j \in A, i_1 \in A^c \setminus B, i_2 \in B,$$

that is, the hypotheses in streams with signal are much easier than in streams with noise, and the hypotheses in B are much harder than in the other streams with

noise. In this case, the first stopping requirement in τ_0 will be easily satisfied, but not the second one, since the streams in B will slow down the growth of the sum of the k_2 least significant negative LLRs.

These observations suggest that the performance of δ_0 can be improved in the above scenario if we essentially "give up" the testing problems in B, presuming that we will make ℓ of the $k_1 - 1$ false positives in these streams. This can be achieved by (i) ignoring the ℓ least significant negative statistics in the second stopping requirement of τ_0, and asking the sum of the *next* k_2 least significant negative statistics to be small upon stopping, and (ii) modifying the decision rule to reject the nulls not only in streams with positive LLR, but also in the ℓ streams with the least significant *negative* LLRs upon stopping. However, if we modify the decision rule in this way, we have spent from the beginning ℓ of the $k_1 - 1$ false positives we are allowed to make. This implies that we need to also modify the first stopping requirement in τ_0 and ask the sum of the $k_1 - \ell$ least significant positive LLRs to be large upon stopping. If we denote by $\widehat{\delta}_\ell := (\widehat{\tau}_\ell, \widehat{D}_\ell)$, the procedure that incorporates the above modifications, then

$$\widehat{\tau}_\ell := \inf\left\{ n \geq 1 : \sum_{j=1}^{k_1-\ell} \widehat{\lambda}^j(n) \geq b \text{ and } \sum_{j=\ell+1}^{\ell+k_2} \check{\lambda}^j(n) \geq a \right\},$$

$$\widehat{D}_\ell := \{\widehat{i}_1(\widehat{\tau}_\ell), \ldots, \widehat{i}_{p(\widehat{\tau}_\ell)}(\widehat{\tau}_\ell)\} \cup \{\check{i}_1(\widehat{\tau}_\ell), \ldots, \check{i}_\ell(\widehat{\tau}_\ell)\},$$

where we omit the dependence on a, b in order to lighten the notation.

By the same token, if there are $\ell < k_2$ streams *with signal* in which the testing problems are much harder than in other streams, it is reasonable to expect that δ_0 may be outperformed by a procedure $\check{\delta}_\ell := (\check{\tau}_\ell, \check{D}_\ell)$, where

$$\check{\tau}_\ell := \inf\left\{ n \geq 1 : \sum_{i=\ell+1}^{\ell+k_1} \widehat{\lambda}^i(n) \geq b \text{ and } \sum_{j=1}^{k_2-\ell} \check{\lambda}^j(n) \geq a \right\},$$

$$\check{D}_\ell := \{\widehat{i}_{\ell+1}(\check{\tau}_\ell), \ldots, \widehat{i}_{p(\check{\tau}_\ell)}(\check{\tau}_\ell)\}.$$

Figure 1 provides a visualization of these stopping rules.

4.2. *The Leap rule*. The previous discussion suggests that the asymmetric Sum-Intersection rule, defined in (23), may be significantly outperformed by some of the procedures, $\{\widehat{\delta}_\ell, 0 \leq \ell < k_1\}$ and $\{\check{\delta}_\ell, 1 \leq \ell < k_2\}$, under some signal configurations, when the multiple testing problem is *asymmetric and/or inhomogeneous*. In this case, we propose combining the above procedures, that is, stop as soon as any of them does so, and use the corresponding decision rule upon stopping. If multiple stopping criteria are satisfied at the same time, we then use the decision rule that rejects the most null hypotheses.

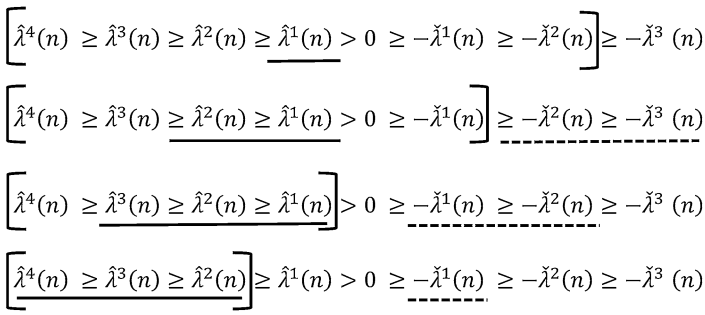


FIG. 1. *Set $J = 7$, $k_1 = 3$, $k_2 = 2$. Suppose at time n, $p(n) = 4$, $q(n) = 3$. Each rule stops when the sum of the terms with solid underline exceeds b, and at the same time the sum of the terms with dashed underline is below $-a$. Upon stopping, the null hypothesis for the streams in the bracket are rejected. Note that by convention (22), $\check{\lambda}^4(n) = \infty$, which makes the stopping rule $\hat{\tau}_2$ have only one condition to satisfy.*

Formally, the proposed procedure $\delta_L := (T_L, D_L)$ is defined as follows:

$$
\begin{aligned}
T_L &:= \min\Big\{ \min_{0\leq \ell < k_1} \widehat{\tau}_\ell, \min_{1 \leq \ell < k_2} \check{\tau}_\ell \Big\}, \\
D_L &:= \Big(\bigcup_{0\leq \ell < k_1, \widehat{\tau}_\ell = T_L} \widehat{D}_\ell \Big) \cup \Big(\bigcup_{1\leq \ell < k_2, \check{\tau}_\ell = T_L} \check{D}_\ell \Big),
\end{aligned} \tag{24}
$$

and we refer to it as "Leap rule," because $\widehat{\delta}_\ell$ (resp., $\check{\delta}_\ell$) "leaps" across the ℓ least significant negative (resp., positive) LLRs.

4.3. *Error control of the Leap rule.* We now show that the Leap rule can control generalized familywise error rates of both types.

THEOREM 4.1. *Assume* (5) *holds. For any $\alpha, \beta \in (0, 1)$ we have that $\delta_L \in \Delta_{k_1,k_2}(\alpha, \beta)$ when the thresholds are selected as follows*:

$$
a = |\log(\beta)| + \log(2^{k_2} C_{k_2}^J), \qquad b = |\log(\alpha)| + \log(2^{k_1} C_{k_1}^J). \tag{25}
$$

PROOF. The proof can be found in Appendix C.1. □

The above threshold values are sufficient for establishing the asymptotic optimality of the Leap rule, but may be conservative in practice. Thus, as in the previous section, we recommend using simulation to find the thresholds that attain the target error probabilities. This means simulating for every $A \subset [J]$ the error probabilities of the Leap rule, $\mathsf{P}_A(|D_L(a, b) \setminus A| \geq k_1)$ and $\mathsf{P}_A(|A \setminus D_L(a, b)| \geq k_2)$, for various pairs of thresholds, a and b, and selecting the values for which the maxima (with respect to A) of the above error probabilities match the nominal levels, α and β, respectively.

As in the previous section, this task is facilitated when the multiple testing problem has a special structure. Specifically, when it is *symmetric* and the user-specified parameters are selected so that $\alpha = \beta$ and $k_1 = k_2$, that is, when condition (12) holds, we can select without any loss of generality the thresholds to be equal ($a = b$). Moreover, if the multiple testing problem is *homogeneous*, the discussion following Theorem 3.1 also applies here.

4.4. *Asymptotic optimality.* For any $B \subset [J]$ and $1 \le \ell \le u \le J$, we denote by

$$\mathcal{I}_1^{(1)}(B) \le \cdots \le \mathcal{I}_1^{(|B|)}(B)$$

the increasingly ordered sequence of $\mathcal{I}_1^j$, $j \in B$, and by

$$\mathcal{I}_0^{(1)}(B) \le \cdots \le \mathcal{I}_0^{(|B|)}(B)$$

the increasingly ordered sequence of $\mathcal{I}_0^j$, $j \in B$, and we set

$$\mathcal{D}_1(B; \ell, u) := \sum_{j=\ell}^{u} \mathcal{I}_1^{(j)}(B) \qquad \text{where } \mathcal{I}_1^{(j)}(B) = \infty \text{ for } j > |B|,$$

$$\mathcal{D}_0(B; \ell, u) := \sum_{j=\ell}^{u} \mathcal{I}_0^{(j)}(B) \qquad \text{where } \mathcal{I}_0^{(j)}(B) = \infty \text{ for } j > |B|.$$

The following lemma provides an asymptotic upper bound on the expected sample size of the stopping times that compose the stopping time of the Leap rule.

LEMMA 4.2. *Assume* (8) *holds. For any* $A \subset [J]$, *we have as* $a, b \to \infty$

$$\mathsf{E}_A[\widehat{\tau}_\ell] \le \max\left\{ \frac{b(1+o(1))}{\mathcal{D}_1(A; 1, k_1 - \ell)}, \frac{a(1+o(1))}{\mathcal{D}_0(A^c; \ell+1, \ell+k_2)} \right\}, \qquad 0 \le \ell < k_1,$$

$$\mathsf{E}_A[\check{\tau}_\ell] \le \max\left\{ \frac{b(1+o(1))}{\mathcal{D}_1(A; \ell+1, \ell+k_1)}, \frac{a(1+o(1))}{\mathcal{D}_0(A^c; 1, k_2 - \ell)} \right\}, \qquad 0 \le \ell < k_2.$$

PROOF. The proof can be found in Appendix C.2. □

If thresholds are selected according to (25), then the upper bounds in the previous lemma are equal (to a first-order asymptotic approximation) to

$$\widehat{L}_A(\ell; \alpha, \beta) := \max\left\{ \frac{|\log \alpha|}{\mathcal{D}_1(A; 1, k_1 - \ell)}, \frac{|\log \beta|}{\mathcal{D}_0(A^c; \ell+1, \ell+k_2)} \right\} \qquad \text{for } \ell < k_1,$$

$$\check{L}_A(\ell; \alpha, \beta) := \max\left\{ \frac{|\log \alpha|}{\mathcal{D}_1(A; \ell+1, \ell+k_1)}, \frac{|\log \beta|}{\mathcal{D}_0(A^c; 1, k_2 - \ell)} \right\} \qquad \text{for } \ell < k_2,$$

and from the definition of Leap rule in (24) it follows that as $\alpha, \beta \to 0$ we have $\mathsf{E}_A[T_L] \le L_A(k_1, k_2, \alpha, \beta)(1 + o(1))$, where

$$L_A(k_1, k_2, \alpha, \beta) := \min\Big\{\min_{0\le \ell < k_1} \widehat{L}_A(\ell; \alpha, \beta), \min_{0\le \ell < k_2} \check{L}_A(\ell; \alpha, \beta)\Big\}. \tag{26}$$

In the next theorem, we show that it is not possible to achieve a smaller ESS, to a first-order asymptotic approximation as $\alpha, \beta \to 0$, proving in this way the asymptotic optimality of the Leap rule.

THEOREM 4.3. *Assume* (8) *holds and that the thresholds in the Leap rule are selected such that* $\delta_L \in \Delta_{k_1,k_2}(\alpha, \beta)$ *and* $a \sim |\log(\beta)|, b \sim |\log(\alpha)|$, *for example, according to* (25). *Then, for any* $A \subset [J]$ *we have as* $\alpha, \beta \to 0$,

$$\mathsf{E}_A[T_L] \sim L_A(k_1, k_2, \alpha, \beta) \sim N_A^*(k_1, k_2, \alpha, \beta).$$

PROOF. In view of the discussion prior to the theorem, it suffices to show that for any $A \subset [J]$ we have as $\alpha, \beta \to 0$ that

$$N_A^*(k_1, k_2, \alpha, \beta) \ge L_A(k_1, k_2, \alpha, \beta)\big(1 - o(1)\big).$$

For the proof of this asymptotic lower bound, we employ similar ideas as in the proof of Theorem 3.2 in the previous section. The change-of-measure argument is more complicated now, due to the interplay of the two kinds of error. We carry out the proof in Appendix C.4. □

REMARK 4.1. When $k_1 = k_2 = 1$, the asymptotic optimality of the Intersection rule was established in [28] only in the i.i.d. case. Since the Leap rule coincides with the Intersection rule when $k_1 = k_2 = 1$, Theorem 4.3 generalizes this result in [28] beyond the i.i.d. case.

We motivated the Leap rule by the inadequacy of the asymmetric Sum-Intersection rule, δ_0, in the case of *asymmetric and/or inhomogeneous* testing problems. In the following corollary, we show that δ_0 is asymptotically optimal when (i) condition (11) holds, which is the case when the multiple testing problem is symmetric and homogeneous, and also (ii) the user-specified parameters are selected in a symmetric way, that is, when (12) holds. In the same setup, we establish the asymptotic optimality of the Intersection rule, δ_I, defined in (10).

COROLLARY 4.4. *Suppose* (8) *and* (11)–(12) *hold and consider the asymmetric Sum-Intersection rule* $\delta_0(b, b)$ *with* $b = b_\alpha$ *and the Intersection rule* $\delta_I(b, b)$ *with* $b = b_\alpha/k_1$, *where* b_a *is defined in* (19) *with* $k = k_1$. *Then* $\delta_0, \delta_I \in \Delta_{k_1,k_1}(\alpha, \alpha)$, *and for any* $A \subset [J]$ *we have as* $\alpha \to 0$ *that*

$$\mathsf{E}_A[\tau_0] \sim \mathsf{E}_A[T_I] \sim \frac{|\log(\alpha)|}{k_1 \mathcal{I}} \sim N_A^*(k_1, k_1, \alpha, \alpha).$$

PROOF. The proof can be found in Appendix C.5. □

REMARK 4.2. In Section 5.2, we will illustrate numerically that when condition (11) is violated, both δ_0 and δ_I fail to be asymptotically optimal.

4.5. *Fixed-sample size rules.* We now focus on the i.i.d. case (15) and consider procedures that stop at a *deterministic* time, which is selected to control the generalized familywise error rates.

For simplicity of presentation, we restrict ourselves to *homogeneous* testing problems, that is, there are densities f_0 and f_1 such that

$$f_0^j = f_0, \qquad f_1^j = f_1 \qquad \text{for every } j \in [J]. \tag{27}$$

This assumption allows us to omit the dependence on the stream index j and write $\mathcal{I}_0 := \mathcal{I}_0^j$, $\mathcal{I}_1 := \mathcal{I}_1^j$ and $\Phi := \Phi^j$, where Φ^j is defined in (17). Moreover, without loss of generality, we apply the MNP rule (14) with the same threshold for each stream.

We further assume that user-specified parameters are selected as follows:

$$k_1 = k_2, \qquad \alpha = \beta^d \qquad \text{for some } d > 0, \tag{28}$$

and that for each $d > 0$ there exists some $h_d \in (-\mathcal{I}_0, \mathcal{I}_1)$ such that

$$\Phi(h_d)/d = \Phi(h_d) - h_d. \tag{29}$$

When $d = 1$, condition (28) reduces to (12) and h_d is equal to 0. However, when $d \neq 1$, we allow for an asymmetric treatment of the two kinds of error.

THEOREM 4.5. *Consider the multiple testing problem* (27) *and assume that the Kullback–Leibler numbers in* (16) *are positive and finite. Further, assume that* (28) *and* (29) *hold. Then as* $\beta \to 0$,

$$\frac{d(1-o(1))}{(2k_1-1)\Phi(h_d)} \le \frac{n^*(k_1,k_1,\beta^d,\beta)}{|\log(\beta)|} \le \frac{\widehat{n}_{\mathrm{NP}}(k_1,k_1,\beta^d,\beta)}{|\log(\beta)|} \sim \frac{d}{k_1\Phi(h_d)}.$$

PROOF. The proof is similar to that of Theorem 3.5, but requires a *generalization* of Chernoff's lemma ([10], Corollary 3.4.6) to account for the asymmetry of the two kinds of error. This generalization is presented in Lemma G.1 and more details can be found in Appendix C.6. □

Theorem 4.5, in conjunction with Theorem 4.3, allows us to quantify the performance loss that is induced by stopping at a deterministic time. Specifically, in the case of testing normal means (Example 2.1), by (18) we have $\mathcal{I} = \mathcal{I}_1 = \mathcal{I}_0$ and for any $d \ge 1$,

$$h_d = \frac{\sqrt{d}-1}{\sqrt{d}+1}\mathcal{I}, \qquad \Phi(h_d) = \frac{d}{(1+\sqrt{d})^2}\mathcal{I}.$$

Thus, by Theorem 4.3 it follows that as $\beta \to 0$,

$$N_A^*(k_1, k_1, \beta^d, \beta) \le \widehat{L}_A(0; \beta^d, \beta) = \begin{cases} \dfrac{|\log(\beta)|}{k_1 \mathcal{I}} & \text{if } |A| < k_1, \\ \dfrac{d|\log(\beta)|}{k_1 \mathcal{I}} & \text{if } |A| \ge k_1. \end{cases}$$

When in particular $d = 1$, that is, $\alpha = \beta$, for any $A \subset [J]$ we have

$$\begin{aligned} 2N_A^*(k_1, k_1, \beta, \beta)\big(1 - o(1)\big) &\le n^*(k_1, k_1, \beta, \beta) \\ &\le \widehat{n}_{\mathrm{NP}}(k_1, k_1, \beta, \beta) \sim 4N_A^*(k_1, k_1, \beta, \beta), \end{aligned}$$

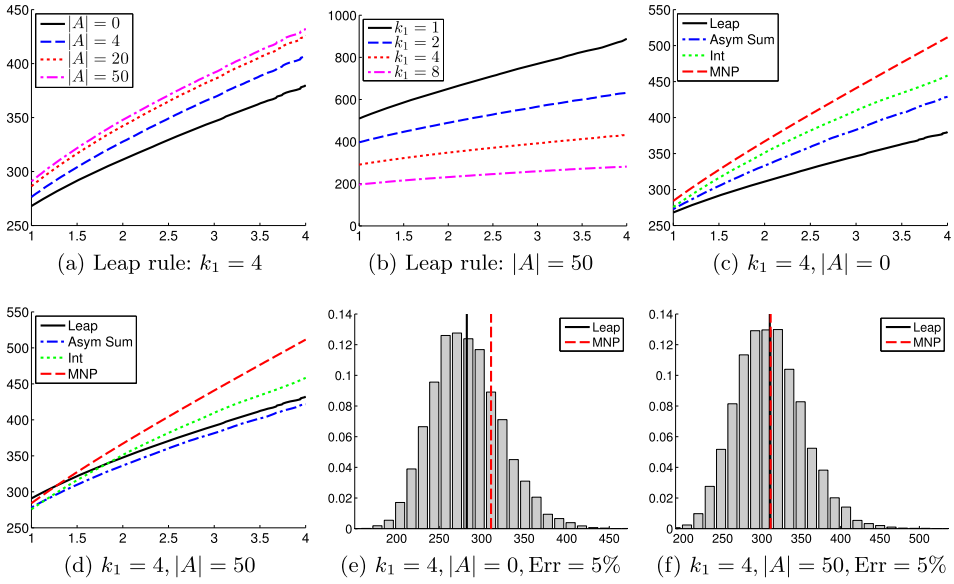
which agrees with the corresponding findings in Section 3.5.

5. Simulations for generalized familywise error rates. In this section, we present two simulation studies that complement our asymptotic optimality theory in Section 4 for procedures that control generalized familywise error rates. In the first study, we compare the Leap rule (24), the Intersection rule (10) and the asymmetric Sum-Intersection rule (23), in a *symmetric and homogeneous* setup where conditions (11) and (12) hold and all three procedures are asymptotically optimal. In the second study, we compare the same procedures when condition (11) is slightly violated, and only the Leap rule enjoys the asymptotic optimality property.

In both studies, we consider the testing of normal means (Example 2.1), with $\sigma_j = 1$ for every $j \in [J]$. This is a *symmetric* multiple testing problem, where the Kullback–Leibler information in the jth testing problem is $\mathcal{I}^j = \mu_j^2/2$. Moreover, we assume that condition (12) holds, that is, $\alpha = \beta$ and $k_1 = k_2$. This implies that we can set the thresholds in each *sequential* procedure to be equal, that is, $a = b$, and as a result the two types of generalized familywise error rates will be the same. Finally, in both studies we include the performance of the fixed-sample size *multiple Neyman–Pearson* (MNP) rule (14), for which the choice of thresholds depends crucially on whether the problem is homogeneous or not.

In what follows, the "error probability (Err)" is the generalized familywise error rate of false positives (3), that is, the *maximum* probability of k_1 false positives, with the maximum taken over all signal configurations. Thus, Err *does not* depend on the true subset of signals $A \subset [J]$.

5.1. *Homogeneous case.* In the first simulation study, we set $\mu_j = 0.25$ for each $j \in [J]$. In this homogeneous setup, the expected sample size (ESS) of all procedures under consideration depend only on the *number* of signals, and we can set the thresholds in the MNP rule, defined in (14), to be equal to 0. Moreover, it suffices to study the performance when the number of signals is no more than $J/2$. We consider $J = 100$ in Figure 2 and $J = 20$ in Figure 3.



(a) Leap rule: $k_1 = 4$ (b) Leap rule: $|A| = 50$ (c) $k_1 = 4, |A| = 0$

(d) $k_1 = 4, |A| = 50$ (e) $k_1 = 4, |A| = 0$, Err $= 5\%$ (f) $k_1 = 4, |A| = 50$, Err $= 5\%$

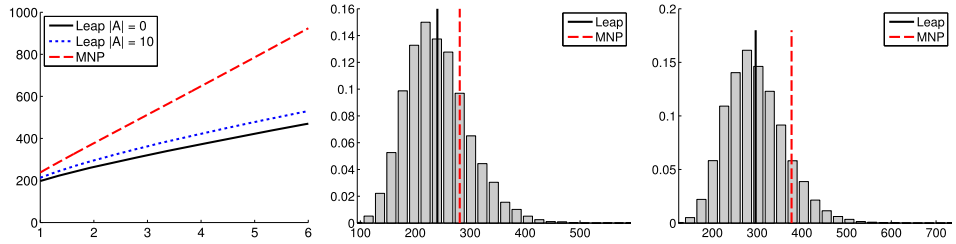
FIG. 2. *Homogeneous case*: $J = 100, k_1 = k_2$. *In* (a)–(d), *the x-axis is* $|\log_{10}(\text{Err})|$ *and the y-axis is the ESS under* P_A. *In* (*e*) *and* (*f*) *are the histogram of the stopping time of the Leap rule with* Err $= 5\%$.

In Figure 2(a), we fix $k_1 = 4$ and evaluate the ESS of the Leap rule for four different cases regarding the number of signals. We see that, for any given Err, the smallest possible ESS is achieved in the boundary case of no signals ($|A| = 0$). This is because some components in the Leap rule only have one condition to be satisfied in the boundary cases (e.g., $\widehat{\tau}_2$ in Figure 1).

In Figure 2(b), we fix the number of signals to be $|A| = 50$ and evaluate the Leap rule for different values of k_1. We observe that there are significant savings in the ESS as k_1 increases and more mistakes are tolerated.

In Figures 2(c) and 2(d), we fix $k_1 = 4$ and compare the four rules for $|A| = 0$ and 50, respectively. In this *symmetric and homogeneous* setup, where (11) and (12) both hold, we have shown that all three sequential procedures are asymptotically optimal. Our simulations suggest that in practice the Leap rule works better when the number of signals, $|A|$, is close to 0 or J, but may perform slightly worse than the asymmetric Sum-Intersection rule, δ_0, when $|A|$ is close to $J/2$.

In Figures 2(c), 2(d) and 3(a), we also compare the performance of the Leap rule with the MNP rule. Further, in Figures 2(e), 2(f), 3(b) and 3(c), we show the histogram of the stopping time of the Leap rule at particular error levels. From these figures, we can see that the best-case scenario for the MNP is when both the number of hypotheses, J, and the error probabilities, Err, are large. Note that this does not contradict our asymptotic analysis, where J is fixed and we let Err go to 0.



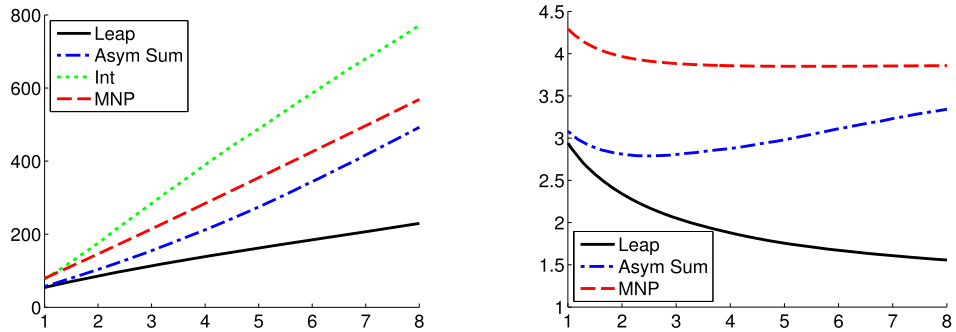
(a) $k_1 = 2$ (b) $k_1 = 2, |A| = 10, \text{Err} = 5\%$ (c) $k_1 = 2, |A| = 10, \text{Err} = 1\%$

FIG. 3. *Homogeneous case*: $J = 20, k_1 = 2$. *In* (a), *the x-axis is* $|\log_{10}(\text{Err})|$ *and the y-axis is the ESS under* P_A. *In* (b) *and* (c) *are the histogram of the stopping time of the Leap rule with* $\text{Err} = 5\%$ *and* 1%.

5.2. *Nonhomogenous case*. In the second simulation study, we set $J = 10$, $\mu_j = 1/6$, $j = 1, 2$, $\mu_j = 1/2$, $j \geq 3$, so that the first two hypotheses are much harder than others. Specifically, $\mathcal{I}^j = 1/72$ for $j = 1, 2$ and $\mathcal{I}^j = 1/8$ for $j \geq 3$.

When the true subset of signals is $A^* = \{6, \ldots, 10\}$, the optimal asymptotic performance, (26), is equal to $8|\log(\text{Err})|$. In Figure 4(a), we plot the ESS against $|\log_{10}(\text{Err})|$, and the ratio of ESS over $8|\log(\text{Err})|$ in Figure 4(b). For the (asymptotically optimal) Leap rule, this ratio tends to 1 as $\alpha \to 0$. In contrast, the other rules have a different "slope" from the Leap rule in Figure 4(a), which indicates that they fail to be asymptotically optimal in this context.

Finally, we note that in such a nonhomogeneous setup, the choice of thresholds for the MNP rule (14) is not obvious. We found that instead of setting $h_j = 0$ for every $j \in [J]$, it is much more efficient to take advantage of the flexibility of generalized familywise error rates, as we did in the construction of the Leap rule in Section 4.2, and set $h_1 = -\infty$, $h_2 = \infty$ and $h_j = 0$ for $j \geq 3$. This choice



(a) ESS under P_{A^*} (b) Normalized by $8|\log(\text{Err})|$

FIG. 4. *Nonhomogeneous case*: $J = 10, k_1 = k_2 = 2, A^* = \{6, \ldots, 10\}$. *The x-axis in both graphs is* $|\log_{10}(\text{Err})|$. *The y-axis in* (a) *is the ESS under* P_{A^*}, *and in* (b) *is the ratio of the ESS over* $8|\log(\text{Err})|$.

"gives up" the first two "difficult" streams by always rejecting the null in the first one and accepting it in the second. The error constraints can then still be met as long as we do not make any mistakes in the remaining "easy" streams. In fact, we see that while the MNP rule behaves significantly worse than the asymptotically optimal Leap rule, it performs better than the Intersection rule, which requires strong evidence from each individual stream in order to stop.

6. Extension to composite hypotheses. We now extend the setup introduced in Section 2, allowing both the null and the alternative hypothesis in each local testing problem to be composite. Thus, for each $j \in [J]$, the distribution of X^j, the sequence of observations in the jth stream, is now parametrized by $\theta^j \in \Theta^j$, where Θ^j is a subset of some Euclidean space, and the hypothesis testing problem in the jth stream becomes

$$\mathsf{H}_0^j : \theta^j \in \Theta_0^j \quad \text{versus} \quad \mathsf{H}_1^j : \theta^j \in \Theta_1^j,$$

where Θ_0^j and Θ_1^j are two disjoint subsets of Θ^j. When $A \subset [J]$ is the subset of streams in which the alternative is correct, we denote by $\boldsymbol{\Theta}_A$ the subset of the parameter space $\boldsymbol{\Theta} := \Theta^1 \times \cdots \times \Theta^J$ that is compatible with A, that is,

$$\boldsymbol{\Theta}_A := \big\{(\theta^1, \ldots, \theta^J) \in \boldsymbol{\Theta} : \theta^i \in \Theta_0^i, \theta^j \in \Theta_1^j \ \forall i \notin A, j \in A\big\}.$$

We denote by $\mathsf{P}_{\theta^j}^j$ the distribution of the jth stream when the value of its local parameter is θ^j. Moreover, we denote by $\mathsf{P}_{A,\theta}$ the underlying probability measure when the subset of signals is A and the parameter is $\theta = (\theta^1, \ldots, \theta^J) \in \boldsymbol{\Theta}_A$, and by $\mathsf{E}_{A,\theta}$ the corresponding expectation. Due to the independence across streams, we have $\mathsf{P}_{A,\theta} = \mathsf{P}_{\theta^1}^1 \otimes \cdots \otimes \mathsf{P}_{\theta^J}^J$.

Our presentation in the case of composite hypotheses will focus on the control of generalized *familywise error rates*; the corresponding treatment of the generalized misclassification rate will be similar. Thus, given $k_1, k_2 \geq 1$ and $\alpha, \beta \in (0, 1)$, the class of procedures of interest now is

$$\Delta_{k_1,k_2}^{\mathrm{comp}}(\alpha, \beta) := \Big\{(T, D) : \max_{A,\theta} \mathsf{P}_{A,\theta}\big(|D \setminus A| \geq k_1\big) \leq \alpha \text{ and } \max_{A,\theta} \mathsf{P}_{A,\theta}\big(|A \setminus D| \geq k_2\big) \leq \beta\Big\},$$

and the goal is the same as the one in Problem 2.2 with $N_A^*(k_1, k_2, \alpha, \beta)$ being replaced by

$$N_{A,\theta}^*(k_1, k_2, \alpha, \beta) := \inf_{(T,D) \in \Delta_{k_1,k_2}^{\mathrm{comp}}(\alpha,\beta)} \mathsf{E}_{A,\theta}[T],$$

and the asymptotic optimality being achieved for every $A \subset [J]$ and $\theta \in \boldsymbol{\Theta}_A$.

6.1. *Leap rule with adaptive log-likelihood ratios.* The proposed procedure in this setup is the Leap rule (24), with the only difference that the local LLR statistics are replaced by statistics that account for the composite nature of the two hypotheses. To be more specific, for every $j \in [J]$ and $n \in \mathbb{N}$ we denote by $\ell^j(n, \theta^j)$ the log-likelihood function (with respect to some σ-finite measure ν_n^j) in the jth stream based on the first n observations, that is,

$$\ell^j(n, \theta^j) := \ell^j(n-1, \theta^j) + \log\bigl(p_{\theta^j}^j(X^j(n)|\mathcal{F}_{n-1}^j)\bigr); \qquad \ell^j(0, \theta^j) := 0,$$

where $p_{\theta^j}^j(X^j(n)|\mathcal{F}_{n-1}^j)$ is the conditional density of $X^j(n)$ given the previous $n-1$ observations in the jth stream. Moreover, for every stream $j \in [J]$ and time $n \in \mathbb{N}$ we denote by $\ell_i^j(n)$ the corresponding *generalized* log-likelihood under H_i^j, that is,

$$\ell_i^j(n) := \sup\{\ell^j(n, \theta^j) : \theta^j \in \Theta_i^j\}, \qquad i = 0, 1.$$

Further, at each $n \in \mathbb{N}$, we select an $\mathcal{F}_n$-measurable estimator of θ, $\widehat{\theta}_n = (\widehat{\theta}_n^1, \dots, \widehat{\theta}_n^J) \in \boldsymbol{\Theta}$, and define the *adaptive log-likelihood* statistic for the jth stream as follows:

$$\ell_*^j(n) := \ell_*^j(n-1) + \log\bigl(p_{\widehat{\theta}_{n-1}^j}^j(X^j(n)|\mathcal{F}_{n-1}^j)\bigr); \qquad \ell_*^j(0) = 0, \tag{30}$$

where $\widehat{\theta}_0 := (\widehat{\theta}_0^1, \dots, \widehat{\theta}_0^J) \in \boldsymbol{\Theta}$ is some deterministic initialization. The proposed procedure in this context is the Leap rule (24), where each LLR statistic $\lambda^j(n)$ is replaced by the following *adaptive* log-likelihood ratio:

$$\lambda_*^j(n) := \begin{cases} \ell_*^j(n) - \ell_0^j(n) & \text{if } \ell_0^j(n) < \ell_1^j(n) \text{ and } \ell_0^j(n) < \ell_*^j(n), \\ -\bigl(\ell_*^j(n) - \ell_1^j(n)\bigr) & \text{if } \ell_1^j(n) < \ell_0^j(n) \text{ and } \ell_1^j(n) < \ell_*^j(n), \\ \text{undefined} & \text{otherwise,} \end{cases} \tag{31}$$

with the understanding that there is no stopping at time n if $\lambda_*^j(n)$ is undefined for some j. Clearly, large positive values of λ_*^j support H_1^j, whereas large negative values of λ_*^j support H_0^j. We denote this modified version of the Leap rule by $\delta_L^*(a, b) = (T_L^*, D_L^*)$.

In the next subsection, we establish the asymptotic optimality of δ_L^* under general conditions. In Appendix D.5, we discuss in more detail the above adaptive statistics, as well as other choices for the local statistics. In Appendix D.4 we demonstrate with a simulation study that if we replace the LLR λ^j by the adaptive statistic λ_*^j (31) in the *Intersection rule* (10) and the *asymmetric Sum-Intersection rule* (23), then these procedures fail to be asymptotically optimal *even in the presence of special structures*. Finally, we should point out that the gains over fixed-sample size procedures are also larger compared to the case of simple hypotheses, as sequential methods are more adaptive to the unknown parameter.

6.2. *Asymptotic optimality.* First of all, for each $j \in [J]$ we generalize condition (7) and assume that for any distinct $\theta^j, \tilde{\theta}^j \in \Theta^j$ there exists a positive number $I^j(\theta^j, \tilde{\theta}^j)$ such that

$$\frac{1}{n}\big(\ell^j(n, \theta^j) - \ell^j(n, \tilde{\theta}^j)\big) \xrightarrow[n\to\infty]{\mathsf{P}^j_{\theta^j} \text{ completely}} I^j(\theta^j, \tilde{\theta}^j). \tag{32}$$

Second, we require that the null and alternative hypotheses in each stream are separated, in the sense that if for each $j \in [J]$ and $\theta^j \in \Theta^j$ we define

$$\mathcal{I}_0^j(\theta^j) := \inf_{\tilde{\theta}^j \in \Theta_1^j} I^j(\theta^j, \tilde{\theta}^j) \quad \text{and} \quad \mathcal{I}_1^j(\theta^j) := \inf_{\tilde{\theta}^j \in \Theta_0^j} I^j(\theta^j, \tilde{\theta}^j), \tag{33}$$

then

$$\mathcal{I}_0^j(\theta^j) > 0 \qquad \forall \theta^j \in \Theta_0^j \quad \text{and} \quad \mathcal{I}_1^j(\theta^j) > 0 \qquad \forall \theta^j \in \Theta_1^j. \tag{34}$$

Finally, we assume that for each $j \in [J]$ and $\varepsilon > 0$,

$$\begin{aligned} &\sum_{n=1}^{\infty} \mathsf{P}^j_{\theta^j}\left(\frac{\ell_*^j(n) - \ell_1^j(n)}{n} - \mathcal{I}_0^j(\theta^j) < -\varepsilon\right) < \infty \qquad \text{for every } \theta^j \in \Theta_0^j, \\ &\sum_{n=1}^{\infty} \mathsf{P}^j_{\theta^j}\left(\frac{\ell_*^j(n) - \ell_0^j(n)}{n} - \mathcal{I}_1^j(\theta^j) < -\varepsilon\right) < \infty \qquad \text{for every } \theta^j \in \Theta_1^j. \end{aligned} \tag{35}$$

We now state the main result of this section, the asymptotic optimality of δ_L^* under the above conditions. The proof is presented in Appendix D.

THEOREM 6.1. *Assume* (32), (34) *and* (35) *hold. Further, assume the thresholds in the Leap rule are selected such that* $\delta_L^*(a, b) \in \Delta^{\text{comp}}_{k_1,k_2}(\alpha, \beta)$ *and* $a \sim |\log(\beta)|$, $b \sim |\log(\alpha)|$, *for example, according to* (25). *Then, for any* $A \subset [J]$ *and* $\theta \in \boldsymbol{\Theta}_A$, *we have as* $\alpha, \beta \to 0$,

$$\mathsf{E}_{A,\theta}[T_L] \sim L_{A,\theta}(k_1, k_2, \alpha, \beta) \sim N^*_{A,\theta}(k_1, k_2, \alpha, \beta),$$

where $L_{A,\theta}(k_1, k_2, \alpha, \beta)$ *is a quantity defined in Appendix D.*1 *that characterizes the asymptotic optimal performance.*

While conditions (32) and (34) are easily satisfied and simple to check, the one-sided complete convergence condition (35) is not as apparent. It is known [32], pages 278–280, that when $\widehat{\theta}_n^j$ is selected to be the maximum likelihood estimator (MLE) of θ^j, condition (35) is satisfied when testing a normal mean with unknown variance, as well as when testing the coefficient of a first-order autoregressive model. In Appendix E, we further show that condition (35) is satisfied when (i) the data in each stream are i.i.d. with some *multiparameter exponential family* distribution, and (ii) the null and the alternative parameter spaces are compact.

7. Conclusion. In this paper, we have considered the sequential multiple testing problem under two error metrics. In the first one, the goal is to control the probability of at least k mistakes, of any kind. In the second one, the goal is to control simultaneously the probabilities of at least k_1 false positives and at least k_2 false negatives. Assuming that the data for the various hypotheses are obtained sequentially in independent streams, we characterized the optimal performance to a first-order asymptotic approximation as the error probabilities vanish, and proposed the first asymptotically optimal procedure for each of the two problems. Procedures that are asymptotically optimal under classical error control ($k = 1$, $k_1 = k_2 = 1$) were found to be suboptimal under *generalized* error metrics. Moreover, in the case of i.i.d. data streams we quantified the asymptotic savings in the expected sample size relative to fixed-sample size procedures.

There are certain questions that remain open. First, we conducted a first-order asymptotic analysis, ignoring higher-order terms in the approximation to the optimal performance. The latter however appears to be nonnegligible in practice [see Figure 4(b)]. Thus, it is an open problem to obtain a more precise characterization of the optimal performance, as well as to examine whether the proposed rules enjoy a stronger optimality property. Second, the number of streams is treated as constant in our asymptotic analysis, but can be very large in practice. It is interesting to consider an enhanced asymptotic regime, where the number of streams also goes to infinity as the error probabilities vanish. Third, although simulation techniques can be used to determine threshold values that guarantee the error control, it is desirable to have closed-form expressions for less conservative threshold values.

Finally, there are several interesting generalizations in various directions. One direction is to relax the assumption that the streams corresponding to the different testing problems are independent. Another direction is to allow for early stopping in some streams, in which case the goal may be to minimize the total number of observations in all streams. Finally, it is interesting to study the corresponding problems with FDR-type error control.

SUPPLEMENTARY MATERIAL

Supplement to "Sequential multiple testing with generalized error control: An asymptotic optimality theory" (DOI: 10.1214/18-AOS1737SUPP; .pdf). In the supplementary file, we present (i) more simulation studies, (ii) proofs of all results in this paper and (iii) additional technical lemmas.

REFERENCES

[1] BARTROFF, J. (2018). Multiple hypothesis tests controlling generalized error rates for sequential data. *Statist. Sinica* **28** 363–398. MR3752265

[2] BARTROFF, J. and LAI, T. L. (2010). Multistage tests of multiple hypotheses. *Comm. Statist. Theory Methods* **39** 1597–1607.

[3] BARTROFF, J. and SONG, J. (2014). Sequential tests of multiple hypotheses controlling type I and II familywise error rates. *J. Statist. Plann. Inference* **153** 100–114. MR3229025

[4] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

[5] BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

[6] BOGDAN, M., CHAKRABARTI, A., FROMMLET, F. and GHOSH, J. K. (2011). Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. *Ann. Statist.* **39** 1551–1579. MR2850212

[7] DE, S. K. and BARON, M. (2012). Sequential Bonferroni methods for multiple hypothesis testing with strong control of family-wise error rates I and II. *Sequential Anal.* **31** 238–262. MR2911288

[8] DE, S. K. and BARON, M. (2012). Step-up and step-down methods for testing multiple hypotheses in sequential experiments. *J. Statist. Plann. Inference* **142** 2059–2070. MR2903412

[9] DE, S. K. and BARON, M. (2015). Sequential tests controlling generalized familywise error rates. *Stat. Methodol.* **23** 88–102. MR3278804

[10] DEMBO, A. and ZEITOUNI, O. (1998). *Large Deviations Techniques and Applications*, 2nd ed. *Applications of Mathematics* (*New York*) **38**. Springer, New York. MR1619036

[11] FORESTI, G. L., REGAZZONI, C. S. and VARSHNEY, P. K. (2003). *Multisensor Surveillance Systems*: *The Fusion Perspective*. Springer, New York.

[12] GUO, W., HE, L. and SARKAR, S. K. (2014). Further results on controlling the false discovery proportion. *Ann. Statist.* **42** 1070–1101. MR3210996

[13] HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6** 65–70. MR0538597

[14] HOMMEL, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75** 383–386.

[15] HOMMEL, G. and HOFFMANN, T. (1988). Controlled uncertainty. In *Multiple Hypothesenprüfung/Multiple Hypotheses Testing* 154–161. Springer, Berlin.

[16] HSU, P. L. and ROBBINS, H. (1947). Complete convergence and the law of large numbers. *Proc. Natl. Acad. Sci. USA* **33** 25–31. MR0019852

[17] KITTUR, A., CHI, E. H. and SUH, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 453–456. ACM, New York.

[18] LEHMANN, E. L. and ROMANO, J. P. (2005). Generalizations of the familywise error rate. *Ann. Statist.* **33** 1138–1154. MR2195631

[19] LEHMANN, E. L., ROMANO, J. P. and SHAFFER, J. P. (2005). On optimality of stepdown and stepup multiple test procedures. *Ann. Statist.* **33** 1084–1108. MR2195629

[20] LI, Y., NITINAWARAT, S. and VEERAVALLI, V. V. (2014). Universal outlier hypothesis testing. *IEEE Trans. Inform. Theory* **60** 4066–4082. MR3225950

[21] MALLOY, M. L. and NOWAK, R. D. (2014). Sequential testing for sparse recovery. *IEEE Trans. Inform. Theory* **60** 7862–7873. MR3285750

[22] MARCUS, R., PERITZ, E. and GABRIEL, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63** 655–660. MR0468056

[23] PEÑA, E. A., HABIGER, J. D. and WU, W. (2011). Power-enhanced multiple decision functions controlling family-wise error and false discovery rates. *Ann. Statist.* **39** 556–583. MR2797856

[24] RAPPAPORT, T. S. (1996). *Wireless Communications*: *Principles and Practice*, 2nd ed. Prentice Hall, Upper Saddle River, NJ.

[25] ROMANO, J. P. and SHAIKH, A. M. (2006). Stepup procedures for control of generalizations of the familywise error rate. *Ann. Statist.* **34** 1850–1873. MR2283720

[26] ROMANO, J. P. and WOLF, M. (2007). Control of generalized error rates in multiple testing. *Ann. Statist.* **35** 1378–1408. MR2351090

[27] SONG, Y. and FELLOURIS, G. (2016). Logarithmically efficient simulation for misclassification probabilities in sequential multiple testing. In 2016 *Winter Simulation Conference* (*WSC*) 314–325.

[28] SONG, Y. and FELLOURIS, G. (2017). Asymptotically optimal, sequential, multiple testing procedures with prior information on the number of signals. *Electron. J. Stat.* **11** 338–363. MR3606774

[29] SONG, Y. and FELLOURIS, G. (2019). Supplement to "Sequential multiple testing with generalized error control: An asymptotic optimality theory." DOI:10.1214/18-AOS1737SUPP.

[30] STOREY, J. D. (2007). The optimal discovery procedure: A new approach to simultaneous significance testing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 347–368. MR2323757

[31] SUN, W. and CAI, T. T. (2009). Large-scale multiple testing under dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 393–424. MR2649603

[32] TARTAKOVSKY, A., NIKIFOROV, I. and BASSEVILLE, M. (2015). *Sequential Analysis: Hypothesis Testing and Changepoint Detection. Monographs on Statistics and Applied Probability* **136**. CRC Press, Boca Raton, FL. MR3241619

[33] TARTAKOVSKY, A. G. (1998). Asymptotic optimality of certain multihypothesis sequential tests: Non-i.i.d. case. *Stat. Inference Stoch. Process.* **1** 265–295. MR2797137

[34] TARTAKOVSKY, A. G., LI, X. R. and YARALOV, G. (2003). Sequential detection of targets in multichannel systems. *IEEE Trans. Inform. Theory* **49** 425–445. MR1966790

[35] WALD, A. (1945). Sequential tests of statistical hypotheses. *Ann. Math. Stat.* **16** 117–186. MR0013275

DEPARTMENT OF STATISTICS
UNIVERSITY OF ILLINOIS,
URBANA–CHAMPAIGN
725 S. WRIGHT STREET
CHAMPAIGN, ILLINOIS 61820
USA
E-MAIL: ysong44@illinois.edu
fexllouri@illinois.edu