Statistica Sinica Preprint No: SS-2025-0042

Title	Sequential Multiple Testing of Multiple Composite Hypotheses: an Asymptotic Optimality Theory with General Information Functions
Manuscript ID	SS-2025-0042
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0042
Complete List of Authors	Yiming Xing
Corresponding Authors	Yiming Xing
E-mails	yimingx4@tongji.edu.cn

Statistica Sinica

SEQUENTIAL MULTIPLE TESTING OF MULTIPLE COMPOSITE HYPOTHESES: AN ASYMPTOTIC OPTIMALITY THEORY WITH GENERAL INFORMATION FUNCTIONS

Yiming Xing

Tongji University

Abstract: This paper considers the problem of testing the marginal distributions of multiple, independent data streams, where for each data stream, multiple composite hypotheses along with an indifference zone are posed. A novel global error metric is proposed, which aims to control the probabilities of making different numbers of misclassifications below different, user-specified levels, and which includes the classical and the generalized misclassification probabilities as special cases. A novel testing procedure is designed and is shown to achieve the minimum expected sample size under all possible distributions, among all tests that control this global error metric below the same levels, asymptotically as any of these levels goes to zero. This asymptotic optimality theory is established allowing temporal dependence and general information functions beyond linear that are considered in most literature. Examples are provided to illustrate the theory and numerical studies are presented to visualize both the asymptotic properties and finite-sample performance.

---

*Author's ORCID ID: 0000-0001-8508-7982

Key words and phrases: asymptotic optimality, multihypothesis testing, multiple testing, non-linear information function, sequential analysis.

1. Introduction

Sequential multiple testing, that is, simultaneously solving multiple hypothesis testing problems based on sequentially sampled data, can be regarded as the abstraction of many real-world problems, such as clinical trials with multiple endpoints Jennison and Turnbull (1999); Bartroff et al. (2012), gene-association studies Zehetmayer et al. (2005); Sarkar et al. (2013), signal/anomaly/intrusion detection from multiple sensors Malloy and Nowak (2014); Cohen and Zhao (2015), etc.

To solve this problem, we need, after sampling every datum, to decide whether or not to continue sampling, and if not, which hypothesis to select for each testing problem. It is clear that the more data one samples, the more information one will have, and the more reliable one's decision can be. However, sampling data incurs a cost. Therefore, trading-off between sampling cost and decision reliability becomes the essence of this problem. To balance this trade-off, there is a Bayesian approach where different weights are assigned to collecting samples and making wrong decisions, and the goal is to minimize the expectation of their weighted sum (see, e.g., Chernoff (1959) in the case of single testing and Hemo et al. (2020) in the case of multiple testing); and there is a frequentists' approach where the goal is to minimize the expected sample size controlling certain metric about making wrong decisions (see, e.g., Wald and Wolfowitz (1948) in the case of single testing and Song and Fellouris (2019) in the case of multiple testing). These two approaches are dual, and we follow the second one in this paper. Specifically, we aim to design a testing procedure that (i) controls the desired global error metric below arbitrary, user-specified levels, and (ii) achieves the minimum expected sample size under every possible distribution among all testing procedures that control the global error metric below the same levels, asymptotically as the levels go to zero.

In the case of familywise error probabilities, that is, the probability of making any false positive error and the probability of making any false negative error, as the global error metric, such an asymptotic optimality theory has been established by many papers with different features that reflect practical situations, e.g., Song and Fellouris (2017); He and Bartroff (2021); Chaudhuri and Fellouris (2024); Xing and Fellouris (2025) with prior information on the number of true alternatives, Cohen and Zhao (2015); Tsopelakos and Fellouris (2023); Xing and Fellouris (2026) with sampling constraints, that is, it is allowed to sample only part of the data streams at a time, and Gafni et al. (2023); Xing et al. (2025) with certain structure in the statistical model. In the case of generalized misclassification probability, that is, the probability of making s or more misclassifications, and in the case of generalized familywise error probabilities, that is, the probability of making s_1 or more false positive errors and the probability of making s_0 or more false negative errors, where $s, s_1, s_0 > 1$ are user-specified integers, as the global error metric, such an asymptotic optimality theory has been established by Song and Fellouris (2019); Tsopelakos and Fellouris (2025) and by Malloy and Nowak (2014); Xing and Fellouris (2023) in a special decentralized setup.

In this paper, we propose a novel global error metric that includes the classical misclassification probability and the generalized misclassification probability as special cases. Specifically, we allow users to assign different tolerance levels for making more than different numbers of misclassifications, typically a higher level for a smaller number and a lower level for a larger number, and aim to control the probability of making more than any number of misclassifications below the corresponding level. For example, when the same level $\alpha \in (0, 1)$ is assigned to making more than $1, 2, \ldots$ misclassifications, we recover the classical misclassification probability; when level one is assigned to making more than $1, \ldots, s-1$ misclassifications and level $\alpha \in (0, 1)$ is assigned to making more than $s, s + 1, \ldots$ misclassifications, we recover the generalized misclassification probability. Furthermore, our global error metric admits numerous, more flexible level assignments, e.g., piecewise flat levels, reflecting that certain numbers of misclassifications trigger certain punishment; linearly, polynomially, or exponentially decaying levels, reflecting that the damage of misclassifications accumulates in a certain way; etc.

Besides, current literature about sequential multiple testing, e.g., those that are cited above, focus on the case where two hypotheses are posed for each testing problem. However, in many scenarios, two hypotheses are inadequate, e.g., when there is an intermediate state between the normal state and the abnormal state, or when we have to distinguish signals into several categories. In such scenarios, it is necessary to pose multiple hypotheses. In the case of sequential single testing, this has been extensively explored, e.g., by Draglia et al. (1999); Nitinawarat et al. (2013); Deshmukh et al. (2021); Xing (2026). However, to the best of the author's knowledge, no existing work considers multiple hypotheses in the case of sequential multiple testing. In this paper, we fill this gap and consider the setup where, for each testing problem, multiple composite hypotheses are posed, and the rest of the hypothesis space is regarded as the indifference zone (for the

---

justification of this term, please see, e.g., Wald (1947)).

In this very general setup and with this very general global error metric, in this paper we propose a novel sequential multiple testing procedure and prove its asymptotic optimality property under quite general conditions. Specifically, the proposed test achieves the minimum expected sample size under every possible distribution (including distributions in the indifference zones) among all testing procedures that control the proposed global error metric below the same levels, to a first-order asymptotic approximation as the minimum of these levels goes to zero. This asymptotic optimality theory holds as long as the likelihood ratio between the true distribution and any alternative hypothesis and the test statistic against that hypothesis satisfy certain strong law of large numbers with respect to the same *information function* (see Assumption 1 and 3). The information function is allowed to increase to infinity in a very general way, which extends the asymptotic optimality theories in most literature that assume a linear information function, e.g., Chapter 5 of Tartakovsky et al. (2014), Song and Fellouris (2019); Xing and Fellouris (2025).

The rest of the paper is organized as follows: In Section 2 we formulate the problem. In Section 3 we establish a universal asymptotic lower bound on the minimum expected sample size. In Section 4 we describe the proposed test, show how to design it to control the global error metric, and establish an asymptotic upper bound on its expected sample size which, combined with the previous asymptotic lower bound, leads to its asymptotic optimality. In Section 5 we present numerical studies and discuss important implementation details of the proposed test. In Section 6 we conclude and pose some future research directions. Examples that feature non-i.i.d. data and/or non-linear information functions, extra numerical studies, and proofs are put in the supplementary material.

Finally, we summarize some notations that will be used throughout this paper. We denote by $\mathbb{N}$ the set of positive integers. For any $M \in \mathbb{N}$, we denote $[M] \equiv \{1, \ldots, M\}$ and $[M]_0 \equiv \{0\} \cup [M]$. For two real numbers x, y, we denote $x \wedge y \equiv \min\{x, y\}$ and $x \vee y \equiv \max\{x, y\}$. For a sequence of real numbers, $\{x_n,\, n \in \mathbb{N}\}$, and a positive real number a that is not necessarily an integer, $x_a \equiv x_{\lceil a \rceil}$. For two sequences of positive real numbers, $\{x_n, n \in \mathbb{N}\}$ and $\{y_n, n \in \mathbb{N}\}$, $x_n \sim y_n$ stands for $\lim_{n \to \infty}(x_n/y_n) = 1$, $x_n \gtrsim y_n$ stands for $\liminf_{n \to \infty}(x_n/y_n) \geq 1$ and $x_n \lesssim y_n$ stands for $\limsup_{n \to \infty}(x_n/y_n) \leq 1$. We use $\log(\cdot)$ to denote the natural logarithm and use $\lg(\cdot)$ to denote the logarithm with base 10. Moreover, we follow the convention that the minimum or infimum over an empty set is $+\infty$. Ties can be broken arbitrarily in all sorting.

---

2. Problem formulation

Consider $K \in \mathbb{N}$ independent data streams,

$$X_k \equiv \{X_k(n),\, n \in \mathbb{N}\}, \quad k \in [K].$$

For each $n \in \mathbb{N}$, denote by $\mathcal{F}_k(n)$ where $k \in [K]$ the σ-algebra generated by the observations from the k_{th} stream up to time n, and by $\mathcal{F}(n)$ the σ-algebra generated by the observations from all streams up to time n, i.e.,

$$\mathcal{F}_k(n) \equiv \sigma\big(X_k(t),\, t \in [n]\big), \quad k \in [K],$$
$$\mathcal{F}(n) \equiv \sigma\big(\mathcal{F}_k(n),\, k \in [K]\big).$$

For each $k \in [K]$, suppose that the distribution of X_k is parametrized by $\theta_k \in \Theta_k$, where Θ_k is some parameter space, denote by P_{k,θ_k} and E_{k,θ_k} the corresponding probability measure and expectation, and we are interested in testing the following $M \geq 2$ hypotheses about θ_k:

$$H_k^m : \theta_k \in \Theta_k^m, \quad m \in [M], \tag{2.1}$$

where $\{\Theta_k^m,\, m \in [M]\}$ are M disjoint subsets of Θ_k. Meanwhile, we denote by $\Theta_k^0 \equiv \Theta_k \backslash \bigcup_{m \in [M]} \Theta_k^m$ the indifference zone, i.e., we are indifferent of selecting any hypothesis when the true parameter value belongs to this zone, which usually happens when it is "intermediate" among two or multiple hypotheses.

Denote the Cartesian product $\boldsymbol{\Theta} \equiv \bigotimes_{k\in[K]} \Theta_k$ as the joint parameter space. For any $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \in \boldsymbol{\Theta}$, denote by $\mathbf{P}_{\boldsymbol{\theta}}$ the joint distribution of all streams when $\boldsymbol{\theta}$ is the concatenation of all local parameter values, i.e., $\mathbf{P}_{\boldsymbol{\theta}} \equiv \bigotimes_{k\in[K]} \mathrm{P}_{k,\theta_k}$, and by $\mathbf{E}_{\boldsymbol{\theta}}$ the corresponding expectation. Moreover, for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, denote by $\boldsymbol{H}_{\boldsymbol{\theta}} \equiv \left(H_{1,\theta_1}, \ldots, H_{K,\theta_K}\right) \in [M]_0^K$ the affiliations of $\boldsymbol{\theta}$, i.e., for each $k \in [K]$, $H_{k,\theta_k} = m \in [M]_0$ if and only if $\theta_k \in \Theta_k^m$.

Our goal is to determine the affiliation of each θ_k, $k \in [K]$ to one of the hypotheses in (2.1) as soon as possible, while controlling the probability of making incorrect decisions. To achieve this, we need to specify a random time indicating when to make these decisions, as well as a decision rule specifying how to make them. Therefore, we call $\delta \equiv (T, \mathbf{D})$ as a *testing procedure*, or in short, a *test* if,

- T is a stopping time with respect to $\{\mathcal{F}(n),\, n \in \mathbb{N}\}$, and
- $\mathbf{D}$ is an $\mathcal{F}(T)$-measurable random vector taking values in $[M]^K$.

The interpretation is that, after taking T samples from every data stream, hypothesis D_k is selected for $k \in [K]$. Note that $D_k = 0$ is not allowed, meaning that we always have to select a hypothesis where we believe θ_k is the most likely to belong. Denote by Δ the family of all tests.

When we make a decision, we are always exposed to the risk of being wrong. For any $s \in [K]$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\delta \in \Delta$, denote by $\alpha_{s,\boldsymbol{\theta}}(\delta)$ the probability that δ makes s or more misclassifications when the true parameter value is $\boldsymbol{\theta}$, i.e.,

$$\alpha_{s,\boldsymbol{\theta}}(\delta) \equiv \mathbf{P}_{\boldsymbol{\theta}}\big(|\boldsymbol{H}_{\boldsymbol{\theta}} \triangle \boldsymbol{D}| \geq s\big), \tag{2.2}$$

where we define

$$\boldsymbol{A} \triangle \boldsymbol{B} \equiv \{k \in [K] : A_k \neq 0,\, A_k \neq B_k\} \text{ for any } \boldsymbol{A} \in [M]_0^K \text{ and } \boldsymbol{B} \in [M]^K.$$

Note that this operator is not symmetric, and when $H_{k,\theta_k} = 0$, regardless of what D_k is, no misclassification is considered to be made for θ_k, reflecting the meaning of "indifference".

Remark 1. When there are two hypotheses and no indifference zone for each testing problem, i.e., when $M = 2$ and $\Theta_k = \Theta_k^1 \cup \Theta_k^2$, $\alpha_{s,\boldsymbol{\theta}}(\cdot)$ is the classical misclassification rate when $s = 1$ (e.g., Malloy and Nowak (2014)) and is the generalized misclassification rate when $2 \leq s \leq K$ (e.g., Song and Fellouris (2019)).

For any $\alpha \in (0, 1]$, denote by $\Delta_s(\alpha)$ the class of tests that control this probability below α for all parameter values in $\boldsymbol{\Theta}$, i.e.,

$$\Delta_s(\alpha) \equiv \{\delta \in \Delta : \alpha_{s,\boldsymbol{\theta}}(\delta) \leq \alpha \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}\},$$

and, for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, denote by $\mathcal{L}_{s,\boldsymbol{\theta}}(\alpha)$ the minimum expected sample size when the true parameter value is $\boldsymbol{\theta}$ among all tests in $\Delta_s(\alpha)$, i.e.,

$$\mathcal{L}_{s,\boldsymbol{\theta}}(\alpha) \equiv \inf \left\{ \mathbf{E}_{\boldsymbol{\theta}}[T] : \delta \in \Delta_s(\alpha) \right\}. \tag{2.3}$$

Furthermore, for any $\boldsymbol{\alpha} \equiv (\alpha_1, \ldots, \alpha_K) \in (0,1]^K$, denote by $\Delta(\boldsymbol{\alpha})$ the class of tests that control $\alpha_{s,\boldsymbol{\theta}}(\delta)$ below α_s for all $s \in [K]$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, i.e.,

$$\Delta(\boldsymbol{\alpha}) \equiv \{\delta \in \Delta : \alpha_{s,\boldsymbol{\theta}}(\delta) \leq \alpha_s \text{ for all } s \in [K] \text{ and } \boldsymbol{\theta} \in \boldsymbol{\Theta}\} \equiv \bigcap_{s \in [K]} \Delta_s(\alpha_s),$$

and, for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, denote by $\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{\alpha})$ the minimum expected sample size when the true parameter value is $\boldsymbol{\theta}$ among all tests in $\Delta(\boldsymbol{\alpha})$, i.e.,

$$\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{\alpha}) \equiv \inf\{\mathrm{E}_{\boldsymbol{\theta}}[T] : \delta \in \Delta(\boldsymbol{\alpha})\}. \tag{2.4}$$

The goal of this paper is to propose a test that,

(i) can be designed to belong to $\Delta(\boldsymbol{\alpha})$ for any $\boldsymbol{\alpha} \in (0,1]^K$ and,

(ii) achieves $\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{\alpha})$ for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ to a first-order asymptotic approximation as $\underline{\boldsymbol{\alpha}} \to 0$,

where we denote $\underline{\boldsymbol{\alpha}} \equiv \min_{k \in [K]} \alpha_k$.

2.1 Notations and assumptions

We fix $k \in [K]$ in this subsection.

We assume that there exists a σ-finite measure ν_k such that $\mathrm{P}_{k,\theta}$ is absolutely continuous with ν_k when restricted to $\mathcal{F}_k(n)$ for any $\theta \in \Theta_k$ and

$n \in \mathbb{N}$, and denote by $\ell_k(n;\theta)$ the logarithm of the corresponding density (Radon-Nikodym derivative), i.e.,

$$\ell_k(n;\theta) \equiv \log \frac{d\mathrm{P}_{k,\theta}}{d\nu_k}(\mathcal{F}_k(n)).$$

We refer to $\ell_k(n;\cdot)$ as the log-likelihood of the local parameter in stream k based on its first n samples.

For any $m \in [M]$ and $\theta \in \Theta_k \backslash \Theta_k^m$, let $\{I_k^m(n;\theta),\, n \in \mathbb{N}\}$ be a sequence of positive real numbers that increases to ∞ and satisfies

$$\lim_{q\in(0,1)\to 1} \lim_{a\to\infty} \frac{I_k^m(qa;\theta)^{-1}}{I_k^m(a;\theta)^{-1}} = 1, \tag{2.5}$$

where, for any $a > 0$, we denote

$$I_k^m(a;\theta)^{-1} \equiv \min\left\{n \in \mathbb{N} : I_k^m(n;\theta) \geq a\right\}.$$

We will avoid using the same notation for reciprocals. For completeness of notations, for any $m \in [M]$ and $\theta \in \Theta_k^m$, let $\{I_k^m(n;\theta),\, n \in \mathbb{N}\}$ be the sequence of zeros.

In order to establish a universal asymptotic lower bound on the optimal expected sample size in (2.3) and (2.4), we make the following assumption regarding the log-likelihoods and the above sequences of real numbers:

Assumption 1. *For any $m \in [M]$, $\theta \in \Theta_k \backslash \Theta_k^m$, and $q \in (0,1)$, there* *exists* $\theta' \in \Theta_k^m$ *so that*

$$\mathrm{P}_{k,\theta}\left(\limsup_{n\to\infty} \frac{\ell_k(n;\theta) - \ell_k(n;\theta')}{I_k^m(n;\theta)} \le \frac{1}{q}\right) = 1. \tag{2.6}$$

Besides, we assume that, for any $\theta \in \Theta_k$,

$$\textit{the sorting of } \{I_k^m(n;\theta),\, m \in [M]\} \textit{ does not change with } n. \tag{2.7}$$

Based on condition (2.7), for any $\theta \in \Theta_k$, we denote by $H'_{k,\theta}, H''_{k,\theta} \in [M]$ the indices corresponding to the smallest and the second smallest values in $\{I_k^m(n;\theta),\, m \in [M]\}$.

Remark 2. (i) Throughout the paper, we refer to $I_k^m(n;\theta)$ as the *information function* between θ and Θ_k^m, which reduces to the *minimum cumulative Kullback-Leibler divergence* between θ and all parameter values in Θ_k^m when X_k are i.i.d. Intuitively, the information functions reflect "cumulative distances" between the true parameter value and the wrong hypotheses. Condition (2.6) imposes upper bounds on the cumulative distances, which will in turn yield a lower bound on the expected time needed to distinguish the true hypothesis from the wrong ones. This lower bound will be formally established in Theorem 1. Condition (2.7) assumes that the sorting of the M cumulative distances is fixed over time, and $H'_{k,\theta}, H''_{k,\theta}$ represent the hypotheses that is the closest and the second closest to θ. Note that when θ belongs to one of the hypotheses, i.e., when $\theta \in \Theta_k^m$ for some $m \in [M]$, then $H_{k,\theta} = H'_{k,\theta} = m$, whereas when θ belongs to the indifference zone, i.e., when $\theta \in \Theta_k^0$, then $H_{k,\theta} = 0$ and $H'_{k,\theta} \in [M]$.

(ii) Note that, as a sequence of real numbers defined by a set of asymptotic properties, $\{I_k^m(n;\theta),\, n \in \mathbb{N}\}$ is not unique. Rather, it should be understood as an equivalence class satisfying these asymptotic properties. Importantly, the asymptotic result it entails, especially the asymptotic approximation to the optimal expected sample size given in Corollary 1, is invariant to the choice of element in the equivalence class.

(iii) In Section S1 of the supplementary material, we study three concrete examples: testing the mean of independent Gaussians with unequal variances, testing the correlation coefficient of an AR(1) series, and testing the transition matrix of a Markov chain. In these examples, the validity of Assumption 1-3 are checked. Specifically, the corresponding information functions take the following form:

$$I_k^m(n;\theta) = \psi_k(n) \inf_{\theta' \in \Theta_k^m} \frac{(\theta-\theta')^2}{2} \quad \text{where } \psi_k(n) \equiv \sum_{t=1}^{n} \frac{1}{\sigma_k(t)^2},$$

$$I_k^m(n;\theta) = n \inf_{\theta' \in \Theta_k^m} \frac{(\theta-\theta')^2}{2(1-\theta^2)},$$

$$I_k^m(n;\Pi) = n \inf_{\Pi' \in \Pi_k^m} \sum_{(i,j)\in[s]^2} \pi_Y(i,j) \log \frac{\Pi(i,j)}{\Pi'(i,j)},$$

where θ in the first and the second examples represent the mean and the correlation coefficient, and Π in the third example represents the transition matrix. For more details, please see Section S1 of the supplementary material.

For any $\theta \in \Theta_k$, we denote by

$$p_{k,\theta}(X_k(n)|\mathcal{F}_k(n-1)) \equiv \exp\left\{\ell_k(n;\theta) - \ell_k(n-1;\theta)\right\}$$

the density of $X_k(n)$ conditional on $\mathcal{F}_k(n-1)$ when the local parameter in the k_{th} stream is θ, where $n \geq 1$ and $\ell_k(0;\theta) = 0$. Denote by $\left\{\hat{\theta}_k(n),\, n \geq 0\right\}$ a sequence of estimators of the local parameter where $\hat{\theta}_k(0)$ is an arbitrary initialization and $\hat{\theta}_k(n)$ is $\mathcal{F}_k(n)$-measurable for $n \geq 1$, and denote

$$\hat{\ell}_k(n) \equiv \hat{\ell}_k(n-1) + \log\left(p_{k,\hat{\theta}_k(n-1)}\left(X_k(n)|\mathcal{F}_k(n-1)\right)\right) \tag{2.8}$$

where $n \geq 1$ and $\hat{\ell}_k(0) \equiv 0$. Moreover, for any $m \in [M]$, denote

$$\ell_k^m(n) \equiv \sup_{\theta \in \Theta_k^m} \ell_k(n;\theta). \tag{2.9}$$

Statistic $\hat{\ell}_k(n)$ is known (see, e.g., (Tartakovsky et al., 2014, Chapter 5)) as the *adaptive log-likelihood* and $\ell_k^m(n)$ as the *generalized log-likelihood* (locally in the m_{th} hypothesis) based on all observations in stream k up to time n.

In order to show that the proposed test terminates almost surely and satisfies the desired error control, we make the following assumption:

Assumption 2. *For any* $m \in [M]$ *and* $\theta \in \Theta_k \backslash \Theta_k^m$,

$$\mathrm{P}_{k,\theta}\left(\lim_{n\to\infty}\left(\hat{\ell}_k(n) - \ell_k^m(n)\right) = \infty\right) = 1. \tag{2.10}$$

Remark 3. We emphasize that this assumption is the only assumption for the validity and reliability of the proposed test, i.e., as long as this assumption holds, the proposed test terminates with probability one and can be designed to control the misclassification probabilities below arbitrary, user-specified levels, as stated in Theorem 2. Intuitively speaking, it requires that the "empirical cumulative distance" between the adaptive log-likelihood and the maximum log-likelihood with respect to any wrong hypothesis increases to infinity with probability one as the sample size accumulates, so every wrong hypothesis will be identified in finite samples. This is a very mild condition and is satisfied in most regular examples with $\hat{\theta}_k(n)$ being the maximum likelihood estimator.

In order to prove an asymptotic upper bound on the expected sample size of the proposed test that matches the universal asymptotic lower bound and, thus, establish a complete asymptotic optimality theory, we make the following assumption:

Assumption 3. *For any* $m \in [M]$, $\theta \in \Theta_k \backslash \Theta_k^m$ *and* $q \in (0, 1)$,

$$\sum_{n=1}^{\infty} \mathrm{P}_{k,\theta}\left(\frac{\hat{\ell}_k(n) - \ell_k^m(n)}{I_k^m(n;\theta)} \leq q\right) < \infty. \tag{2.11}$$

Remark 4. (i) While condition (2.10) requires the cumulative empirical distance to increase to infinity almost surely, condition (2.11) further imposes a lower bound on its "increasing speed", which in turn suffices for deriving an upper bound on the time needed to reach any pre-specified threshold, as formalized in Theorem 3.

(ii) It is shown in (Song and Fellouris, 2019, Section E) that all Assumptions 1-3 are satisfied with $I_k^m(\cdot;\theta)$ being linear, i.e., $I_k^m(n;\theta) = nI_{k,\theta}^m$ for some constant $I_{k,\theta}^m > 0$, when for each data stream, the data are i.i.d. whose distribution belongs to some multi-parameter exponential family, the parameter space corresponding to each hypothesis is compact, and the maximum likelihood estimator is used as the adaptive estimator. In Section S1 of the supplementary material, we present various examples where all assumptions are satisfied but the data are not i.i.d. and the information functions are not necessarily linear.

3. Universal asymptotic lower bound

In this section, we establish an asymptotic lower bound on (2.3) and (2.4) as $\underline{\boldsymbol{\alpha}} \to 0$. Prior to this, we introduce some notations. For any $s \in [K]$ and

$\boldsymbol{\theta} \in \boldsymbol{\Theta}$, denote

$$\boldsymbol{I}_s(n;\boldsymbol{\theta}) \equiv \min_{\substack{S \subseteq [K] \\ |S|=s}} \sum_{k \in S} I_k^{H''_{k,\theta_k}}(n;\theta_k), \tag{3.12}$$

i.e., the sum of the s smallest ones in $\left\{ I_k^{H''_{k,\theta_k}}(n;\theta_k),\ k \in [K] \right\}$. Note that under condition (2.7) the identities of the s smallest ones are fixed over time. In addition, for any $a > 0$, denote

$$\boldsymbol{I}_s(a;\boldsymbol{\theta})^{-1} \equiv \min \left\{ n \in \mathbb{N} : \boldsymbol{I}_s(n;\boldsymbol{\theta}) \geq a \right\}.$$

Theorem 1. *Suppose that Assumption 1 holds. For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,*

$$\mathcal{L}_{s,\boldsymbol{\theta}}(\alpha) \gtrsim \boldsymbol{I}_s(|\log \alpha|;\boldsymbol{\theta})^{-1} \text{ for any } s \in [K] \text{ as } \alpha \to 0, \tag{3.13}$$

and, thus,

$$\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{\alpha}) \geq \max_{s \in [K]} \mathcal{L}_{s,\boldsymbol{\theta}}(\alpha_s) \gtrsim \max_{s \in [K]} \boldsymbol{I}_s(|\log \alpha_s|;\boldsymbol{\theta})^{-1} \text{ as } \underline{\boldsymbol{\alpha}} \to 0. \tag{3.14}$$

4. The proposed test

In this section, we describe the proposed test, show how to select its thresholds to control the misclassification probabilities below user-specified levels, and establish an asymptotic upper bound on its expected sample size that matches the universal asymptotic lower bound in the previous section.

Algorithm 1 The proposed test

```
Input positive thresholds $a_1, \ldots, a_K$.
Initialize adaptive log-likelihood $\hat{\ell}_k = 0$ and adaptive estimator $\hat{\theta}_k \in \Theta_k$ for $k \in [K]$
Initialize $\xi_k = 0$ for $k \in [K]$ and sort them as $\xi_{(1)} \leq \cdots \leq \xi_{(K)}$
while there exists $1 \leq s \leq K$ so that $\sum_{k=1}^{s} \xi_{(k)}(n) < a_s$, do
    $n = n + 1$
    for $1 \leq k \leq K$ do
        sample new $X_k$
        update adaptive log-likelihood $\hat{\ell}_k$ as in (2.8)
        update adaptive estimator $\hat{\theta}_k$ and local MLEs $\{\hat{\theta}_k^m, m \in [M]\}$
        update local generalized log-likelihoods $\{\ell_k^m, m \in [M]\}$ as in (2.9)
        sort $\{\ell_k^m, m \in [M]\}$ as $\ell_k^{(1)} \leq \cdots \leq \ell_k^{(M)}$
        compute $\xi_k = \hat{\ell}_k - \ell_k^{(M-1)}$
    end for
    sort $\{\xi_k, k \in [K]\}$ as $\xi_{(1)} \leq \cdots \leq \xi_{(K)}$
end while
Output sample size $\hat{T} = n$ and decision $\hat{D}_k = \operatorname{argmax}_{m \in [M]} \ell_k^m$ for $k \in [K]$
```

4.1 Description

First of all, we introduce some notations. For every $k \in [K]$ and $n \in \mathbb{N}$, we denote the order statistics of $\{\ell_k^m(n),\, m \in [M]\}$ as

$$\ell_k^{(1)}(n) \leq \cdots \leq \ell_k^{(M)}(n),$$

and denote the gap between $\hat{\ell}_k(n)$ and the second largest of these order statistics as $\xi_k(n)$, i.e.,

$$\xi_k(n) \equiv \hat{\ell}_k(n) - \ell_k^{(M-1)}(n).$$

Besides, for every $n \in \mathbb{N}$, we denote the order statistics of $\{\xi_k(n),\, k \in [K]\}$ as

$$\xi_{(1)}(n) \leq \cdots \leq \xi_{(K)}(n).$$

For some threshold $\boldsymbol{a} = (a_1, \ldots, a_K) \in (0, \infty)^K$ to be determined, the proposed test is defined as follows:

$$\begin{aligned} \hat{T} &\equiv \inf \left\{ n \geq 1 : \sum_{k=1}^{s} \xi_{(k)}(n) \geq a_s \text{ for all } s \in [K] \right\}, \\ \hat{D}_k &\equiv \operatorname*{argmax}_{m \in [M]} \ell_k^m(\hat{T}), \quad \forall\, k \in [K]. \end{aligned} \tag{4.15}$$

Suppressing the dependence on $\boldsymbol{a}$, we denote it by $\hat{\delta} = (\hat{T}, \hat{\boldsymbol{D}})$.

Briefly speaking, the proposed test samples all streams until for any $s \in [K]$, the sum of the s smallest ones in $\{\xi_k,\, k \in [K]\}$ is at least a_s, at which time the hypothesis corresponding to the largest one in $\{\ell_k^m,\, m \in [M]\}$ is selected for every stream $k \in [K]$. An algorithmic description is presented in Algorithm 1.

Remark 5. (i) Here are some intuitions: (i.i) ξ_k reflects the evidence against all but one hypothesis, i.e., $\underset{m \in [M]}{\operatorname{argmax}}\ \ell_k^m$, so the larger its value, the more evidence in favor of this hypothesis. (i.ii) In order to control the probability of making one misclassification, we need to sample until any of $\{\xi_k,\, k \in [K]\}$ is large enough. In order to control the probability of making $s \geq 2$ misclassifications, this criterion can be relaxed and we need to sample until any s-sum of $\{\xi_k,\, k \in [K]\}$ is large enough. Universal values of how large they need to be in order to control the misclassification probabilities below desired levels are provided in the next subsection.

(ii) The proposed test is scalable with respect to dimension (the number of data streams). To see this, note that its implementation involves running one adaptive multi-hypothesis sequential probability ratio test (which is the state-of-the-art solution for single-stream multi-hypothesis testing, see, e.g., (Tartakovsky et al., 2014, Chapter 5)) in each stream, and sorting and thresholding the statistics across streams at every time instant. Thus, the total computational complexity is $O(K \log K)$ times that required for solving the testing problem in one stream.

4.2 Error control

The next theorem establishes upper bounds on the misclassification probabilities for any given thresholds, which in turn imply a selection of the thresholds that ensure the misclassification probabilities below arbitrary desired levels.

Theorem 2. *Suppose that Assumption 2 holds. For any* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *and* $\boldsymbol{a} \in (0, \infty)^K$,

$$\mathbf{P}_{\boldsymbol{\theta}}(\hat{T} < \infty) = 1, \tag{4.16}$$

and

$$\alpha_{s,\boldsymbol{\theta}}(\hat{\delta}) \leq \binom{K}{s}(M-1)^s e^{-a_s} \text{ for every } s \in [K]. \tag{4.17}$$

Thus, for any $\boldsymbol{\alpha} \in (0,1]^K$, $\hat{\delta} \in \Delta(\boldsymbol{\alpha})$ *if*

$$a_s = |\log \alpha_s| + \log\left(\binom{K}{s}(M-1)^s\right) \text{ for every } s \in [K]. \tag{4.18}$$

Remark 6. This selection of thresholds is universal to all underlying distributions. Thus, it can be very conservative without utilizing any problem-specific information. For practical use, we recommend doing Monte-Carlo simulations offline to find thresholds that approximately equate the actual misclassification probabilities with the desired levels. Since the misclassification probabilities are very small, plain Monte-Carlo is very inefficient, and importance sampling (see, e.g., Siegmund (1976); Bucklew (2010)) is recommended. For the selection of the importance sampling distribution, please see the numerical studies in Section 5.

4.3 Asymptotic optimality

The next theorem establishes an asymptotic upper bound on the expected sample size of the proposed test, which, combined with the universal asymptotic lower bound in Theorem 1, implies its asymptotic optimality property.

Theorem 3. *Suppose that Assumption 3 holds. For any* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$*,*

$$\mathbf{E}_{\boldsymbol{\theta}}[\hat{T}] \lesssim \max_{s \in [K]} \boldsymbol{I}_s(a_s; \boldsymbol{\theta})^{-1} \;\; \textit{as } \overline{\boldsymbol{a}} \to \infty, \tag{4.19}$$

where we denote $\overline{\boldsymbol{a}} \equiv \max_{k \in [K]} a_k$.

Combining Theorem 1-3, we have the following corollary.

Corollary 1. *Suppose that Assumption 1-3 hold. If* $\hat{\delta}$ *is designed so that* $\hat{\delta} \in \Delta(\boldsymbol{\alpha})$ *for any* $\boldsymbol{\alpha} \in (0,1]^K$ *and* $a_s \sim |\log \alpha_s|$ *for any* $s \in [K]$ *as* $\underline{\boldsymbol{\alpha}} \to 0$*, e.g., as in* (4.18)*, then, for any* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$*,*

$$\mathbf{E}_{\boldsymbol{\theta}}[\hat{T}] \sim \mathcal{L}_{\boldsymbol{\theta}}(\alpha) \sim \max_{s \in [K]} \boldsymbol{I}_s(|\log \alpha_s|; \boldsymbol{\theta})^{-1} \;\; \textit{as } \underline{\boldsymbol{\alpha}} \to 0. \tag{4.20}$$

5. Numerical studies and implementation details

In this section, we present a numerical study of testing the mean of i.i.d. Gaussian sequences. The goal is to illustrate some details in the implemen-

---

tation of the proposed test, including using a pilot stage to obtain better initial estimates of the parameters and using importance-sampling based Monte-Carlo simulations to find sharper thresholds, and to visualize its finite-sample performance and asymptotic property. An extra numerical study of testing the correlation coefficient of AR(1) series is presented in Section S1 of the supplementary material.

Specifically, consider K i.i.d. Gaussian sequences with unknown mean θ_k and unit variance, i.e., $N(\theta_k, 1)$, for $k \in [K]$. For each of them, consider the following testing problem:

$$\Theta_k^1 = [-1, -0.5], \quad \Theta_k^2 = [-0.25, 0.25], \quad \Theta_k^3 = [0.5, 1],$$
$$\Theta_k^0 = (-0.5, -0.25) \cup (0.25, 0.5),$$

that is, $M = 3$ and $\Theta_k = [-1, 1]$.

For any $k \in [K]$, using $N(0, 1)$ as the reference measure, we have

$$\ell_k(n; \theta_k) = \left(\theta_k \bar{X}_k(n) - \frac{1}{2}\theta_k^2\right) n, \quad \text{where } \bar{X}_k(n) \equiv \frac{1}{n}\sum_{t=1}^{n} X_k(t).$$

It has been checked in (Song and Fellouris, 2019, Section E) that Assumption 1-3 hold with

$$I_k^m(n; \theta_k) = n \inf_{\theta_k' \in \Theta_k^m} \frac{n}{2}(\theta_k - \theta_k')^2 \quad \text{for } m \in [M].$$

The global Maximum Likelihood Estimator (MLE) is

$$\hat{\theta}_k(n) = \begin{cases} \text{arbitrary in } [-1,1] & \text{for } n = 0, \\ (\bar{X}_k(n) \wedge 1) \vee -1 & \text{for } n \geq 1, \end{cases}$$

and the adaptive log-likelihood in (2.8) is

$$\hat{\ell}_k(n) = \sum_{t=1}^{n} \left(\hat{\theta}_k(t-1) X_k(t) - \frac{1}{2} \hat{\theta}_k(t-1)^2 \right), \quad n \geq 1. \tag{5.21}$$

Besides, the local MLEs are

$$\begin{aligned} \hat{\theta}_k^1(n) &= (\bar{X}_k(n) \wedge -0.5) \vee -1, \\ \hat{\theta}_k^2(n) &= (\bar{X}_k(n) \wedge 0.25) \vee -0.25, \quad \text{for } n \geq 1, \\ \hat{\theta}_k^3(n) &= (\bar{X}_k(n) \wedge 1) \vee 0.5, \end{aligned}$$

and the local generalized log-likelihoods in (2.9) is

$$\ell_k^m(n) = \ell_k(n; \hat{\theta}_k^m(n)) = n \left(\hat{\theta}_k^m(n) \bar{X}_k(n) - \frac{1}{2} \hat{\theta}_k^m(n)^2 \right) \text{ for } m \in [M] \text{ and } n \geq 1.$$

5.1 Implementation details

Although suffices for the theoretical analysis, an arbitrary initialization and the first a few noisy estimators of $\hat{\theta}_k(n)$ can cause $\hat{\ell}_k(n)$ to take long to stabilize. To alleviate this problem, we recommend using a pilot stage to obtain a good initial estimate of the unknown parameter. Specifically, let $X_k(-n_0), \ldots, X_k(-1)$ be $n_0 \in \mathbb{N}$ observations that are also i.i.d. following

$N(\theta_k, 1)$, and define the global MLE as

$$\hat{\theta}_k(n) \equiv \left(\frac{\sum_{t=-n_0}^{-1} X_k(t) + \sum_{t=1}^{n} X_k(t)}{n_0 + n} \wedge 1 \right) \vee -1 \text{ for } n \geq 0.$$

The adaptive log-likelihood is the same as in (5.21) with this new $\hat{\theta}_k(n)$ plugged in. In this numerical study, we use $n_0 = 10$.

As stated in the remark after Theorem 2, the universal selection of thresholds in (4.18) can be very conservative due to ignorance of any problem-specific information. In order to find sharp thresholds, we need to estimate the actual misclassification probabilities under different thresholds, and find the ones whose actual misclassification probabilities approximately equate the desired levels. Specifically, for any $\boldsymbol{a} \in (0, \infty)^K$ and $s \in [K]$, we need to evaluate

$$\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \alpha_{s,\boldsymbol{\theta}}(\hat{\delta}) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbf{P}_{\boldsymbol{\theta}} \left(|\boldsymbol{H}_{\boldsymbol{\theta}} \triangle \hat{\boldsymbol{D}}| \geq s \right).$$

Due to stochastic monotonicity of the Gaussian distribution and symmetry of the parameter space, it suffices to evaluate this quantity for $\boldsymbol{\theta}$ so that $\theta_k \in \{-0.5, -0.25\}$ for all $k \in [K]$. For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\alpha_{s,\boldsymbol{\theta}}(\hat{\delta})$ is very small when a_s is large, so it is very inefficient to evaluate it directly via plain Monte-Carlo. Therefore, we apply importance sampling (see, e.g., Siegmund (1976); Bucklew (2010)).

Specifically, to evaluate $\mathbf{P}_{\boldsymbol{\theta}}(\Gamma)$ where $\Gamma \in \mathcal{F}(\hat{T})$, the key is to find another probability distribution, $\mathbf{P}^*_{\boldsymbol{\theta}}$, (i) that is similar to $\mathbf{P}_{\boldsymbol{\theta}}$, but (ii) under which Γ occurs with high probability. Then, based on the following Wald's likelihood ratio identity:

$$\mathbf{P}_{\boldsymbol{\theta}}(\Gamma) = \mathbf{E}_{\boldsymbol{\theta}}\left[1\{\Gamma\}\right] = \mathbf{E}^*_{\boldsymbol{\theta}}\left[\left(\frac{d\mathbf{P}^*_{\boldsymbol{\theta}}}{d\mathbf{P}_{\boldsymbol{\theta}}}(\mathcal{F}(T))\right)^{-1} 1\{\Gamma\}\right],$$

one can estimate $\mathbf{P}_{\boldsymbol{\theta}}(\Gamma)$ by simulating $\left(\frac{d\mathbf{P}^*_{\boldsymbol{\theta}}}{d\mathbf{P}_{\boldsymbol{\theta}}}(\mathcal{F}(T))\right)^{-1} 1\{\Gamma\}$ under $\mathbf{P}^*_{\boldsymbol{\theta}}$ for numerous rounds and taking the average.

Based on experiments, to estimate $\alpha_{s,\boldsymbol{\theta}}(\hat{\delta})$ we recommend using the following importance sampling distribution:

$$\mathbf{P}^*_{\boldsymbol{\theta}} = \frac{1}{\binom{K}{s}} \sum_{S\subseteq[K],\,|S|=s} \bigotimes_{k\in S} \mathrm{P}_{k,C(\theta_k)} \bigotimes_{k\in[K]\setminus S} \mathrm{P}_{k,\theta_k},$$

where $C(\theta_k)$ represents the parameter value in $\Theta_k^{H''_{k,\theta_k}}$ that is the closest to θ_k. That is to say, we randomly select s out of the K data streams and change their local parameters to the closest ones in the wrong hypotheses. The intuition is that (i) under this distribution, making at least m misclassifications with respect to $\boldsymbol{\theta}$ is very likely, whose probability can be shown to be at least $1-O(e^{-a_s})$, and (ii) it is close to $\mathbf{P}_{\boldsymbol{\theta}}$ provided the requirement in (i).

5.2 Results

In this subsection, we present results of numerical studies. Specifically, we consider $K = 4$ data streams and we evaluate the probabilities of making

$s \in [K]$ or more misclassifications, which we denote by

$$\mathrm{Err}_s \equiv \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \alpha_{s,\boldsymbol{\theta}}(\hat{\delta}), \quad s \in [K], \tag{5.22}$$

and the expected sample size under $\boldsymbol{\theta}^* = (-0.5, -0.25, 0.25, 0.5)$, which we denote by

$$\mathrm{ESS}^* \equiv \mathrm{E}_{\boldsymbol{\theta}^*}[\hat{T}], \tag{5.23}$$

of the proposed test for $\boldsymbol{a} = (a, 2a, 3a, 4a)$ and $a \in \{1, 2, \ldots, 20\}$. All simulations are based on 10^5 rounds.

We first check the efficiency of our importance-sampling-based Monte-Carlo estimates of the misclassification probabilities. In Figure 1 we plot the relative error, i.e., standard deviation of the estimate based on 10^5 simulation rounds divided by the true value, of Err_s against $-\lg(\mathrm{Err}_s)$ for $1 \leq s \leq K$. We can see all relative errors are below 5%, reflecting high estimation accuracy. Note that if plain Monte-Carlo is used to estimate the probability of an event whose true probability is p, then the relative error based on n simulation rounds would be

$$\frac{\mathrm{sd}(\hat{p}_n)}{p} = \frac{\sqrt{p(1-p)/n}}{p} \approx \frac{1}{\sqrt{np}},$$

so the number of simulation rounds needed to let the relative error below 5% would be $400/p$, which is huge for small p and unacceptable for p as small as 10^{-40}.

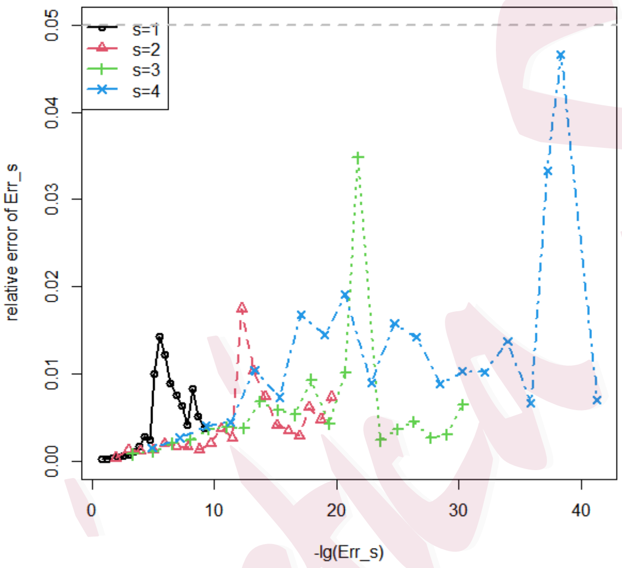

Figure 1: Relative errors of the importance-sampling-based Monte-Carlo estimates of the misclassification probabilities versus $-\lg$ of the estimates themselves. Each estimate is based on 10^5 simulation rounds.

Next, we compare the universal selection of thresholds in (4.18) with the sharp selection of thresholds from Monte-Carlo simulations, where the latter is the set of thresholds whose actual misclassification probabilities hit the target levels. This is plotted in Figure 2. We can see that there is basically a constant gap between the universal thresholds and the sharp thresholds, which does not matter in the asymptotic analysis but will have an impact in real applications. Therefore, we recommend doing Monte-Carlo simulations offline to find sharp thresholds before real implementation of the procedure, especially when the target misclassification probabilities are not very small. As discussed in Song and Fellouris (2017); Xing and Fellouris (2024), this step is unavoidable, unless strong distributional assumptions are imposed and a tailored threshold selection strategy is designed accordingly.

Finally, we corroborate our asymptotic theory via an experiment. According to the asymptotic approximation to the optimal expected sample size in (4.20), with the choice of thresholds we use, $\boldsymbol{a} = (a, 2a, 3a, 4a)$, all of

$$\boldsymbol{I}_s(|\log(\mathrm{Err}_s)|; \boldsymbol{\theta}^*)^{-1}, \quad s \in [K] \tag{5.24}$$

should be asymptotic approximations to ESS^*. This claim is supported by Figure 3.

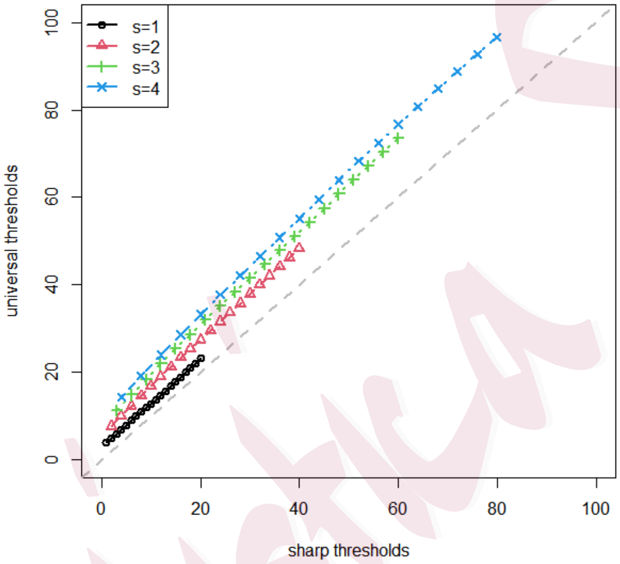

Figure 2: Universal thresholds in (4.18) versus sharp thresholds obtained from Monte-Carlo simulations.

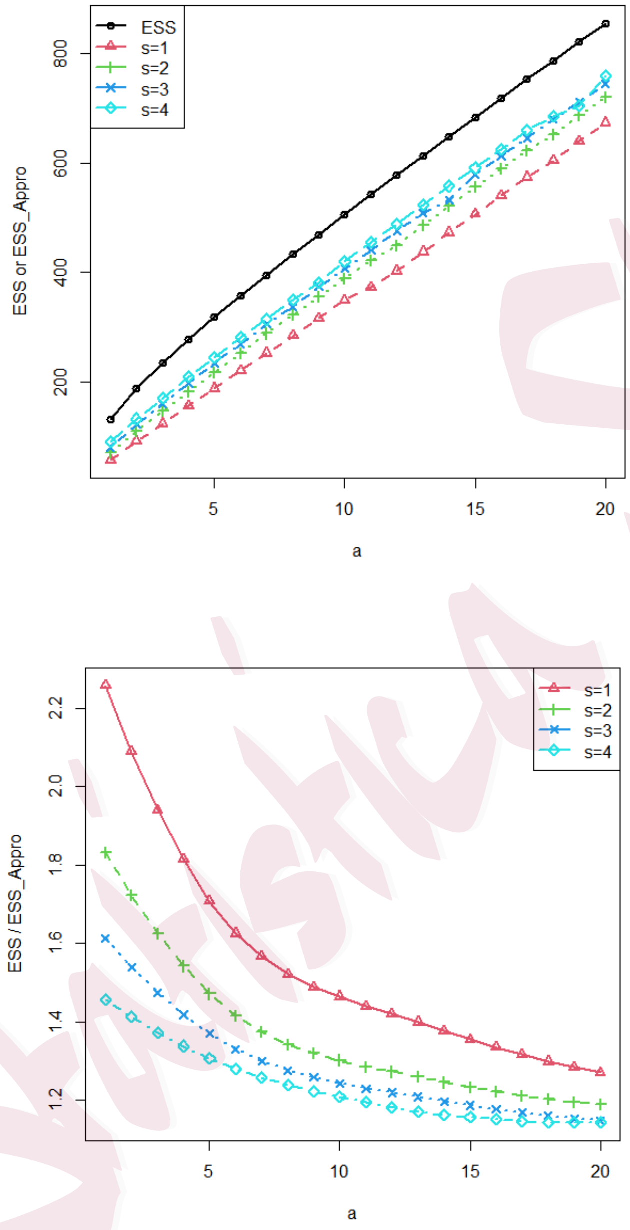

Figure 3: Expected sample size of the proposed test in (5.23) and its asymptotic approximations in (5.24) (top), and their ratios (bottom), against $a \in \{1, 2, \dots, 20\}$ when thresholds $\boldsymbol{a} = \{a, 2a, 3a, 4a\}$. All curves at the bottom converge to one as a increases.

6. Conclusion

In this paper, we expand the scope of sequential multiple testing by incorporating multiple composite hypotheses and an indifference zone for each testing problem. A novel global error metric is proposed that provides practitioners with flexibility to set distinct tolerance levels for different numbers of misclassifications. A new testing procedure is designed that controls this global error metric below arbitrary, user-specified levels, and is shown to achieve the minimum expected sample size asymptotically as the levels go to zero. This asymptotic optimality theory is established allowing general information functions beyond linear that is considered in most literature, and is verified in various examples and numerical studies.

Here are some directions that deserve further exploration: (i) It is straightforward to follow how the misclassification probability is generalized in this paper to generalize the familywise error probabilities, but it is highly non-trivial to establish a similar asymptotic optimality theory in the case of multiple hypotheses. (ii) Derive a high-order asymptotic approximation for the expected sample size of the proposed test under stronger distributional assumptions. (iii) Allow dependence across streams, following (Chaudhuri and Fellouris, 2024), while keeping the computational complexity low. (iv) Allow making decisions at different times, in the spirit of Xing and Fellouris

(2025).

Supplementary material

In the supplementary material, we illustrate the general theory through three concrete examples, present extra numerical studies of testing the correlation coefficient of autoregressive data, and present all proofs and a discussion about model misspecification.

Acknowledgements

The author would like to thank the Editor, the Associate Editor and the reviewers for valuable comments and constructive suggestions, which have greatly improved the paper. The author is supported by National Natural Science Foundation of China (No. `12501379`), Shanghai Rising-Star Program (No. `24YF2748500`), and Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science (East China Normal University), Ministry of Education (No. `KLATASDS2501`).

References

Bartroff, J., T. L. Lai, and M.-C. Shih (2012). *Sequential experimentation in clinical trials: design and analysis*, Volume 298. Springer Science &

Business Media.

Bucklew, J. (2010). *Introduction to Rare Event Simulation* (1st ed.). Springer Publishing Company, Incorporated.

Chaudhuri, A. and G. Fellouris (2024). Joint sequential detection and isolation for dependent data streams. *The Annals of Statistics 52*(5), 1899–1926.

Chernoff, H. (1959). Sequential design of experiments. *The Annals of Mathematical Statistics 30*(3), 755–770.

Cohen, K. and Q. Zhao (2015). Asymptotically optimal anomaly detection via sequential testing. *IEEE Transactions on Signal Processing 63*(11), 2929–2941.

Deshmukh, A., V. V. Veeravalli, and S. Bhashyam (2021). Sequential controlled sensing for composite multihypothesis testing. *Sequential Analysis 40*(2), 259–289.

Draglia, V., A. Tartakovsky, and V. Veeravalli (1999). Multihypothesis sequential probability ratio tests .i. asymptotic optimality. *IEEE Transactions on Information Theory 45*(7), 2448–2461.

Gafni, T., B. Wolff, G. Revach, N. Shlezinger, and K. Cohen (2023).

Anomaly search over discrete composite hypotheses in hierarchical statistical models. *IEEE Transactions on Signal Processing 71*, 202–217.

He, X. and J. Bartroff (2021). Asymptotically optimal sequential fdr and pfdr control with (or without) prior information on the number of signals. *Journal of Statistical Planning and Inference 210*, 87–99.

Hemo, B., T. Gafni, K. Cohen, and Q. Zhao (2020). Searching for anomalies over composite hypotheses. *IEEE Transactions on Signal Processing 68*, 1181–1196.

Jennison, C. and B. W. Turnbull (1999). *Group sequential methods with applications to clinical trials.* CRC Press.

Malloy, M. L. and R. D. Nowak (2014). Sequential testing for sparse recovery. *IEEE Transactions on Information Theory 60*(12), 7862–7873.

Nitinawarat, S., G. K. Atia, and V. V. Veeravalli (2013). Controlled sensing for multihypothesis testing. *IEEE Transactions on automatic control 58*(10), 2451–2464.

Sarkar, S. K., J. Chen, and W. Guo (2013). Multiple testing in a two-stage adaptive design with combination tests controlling fdr. *Journal of the American Statistical Association 108*(504), 1385–1401.

Siegmund, D. (1976). Importance Sampling in the Monte Carlo Study of Sequential Tests. *The Annals of Statistics 4*(4), 673 – 684.

Song, Y. and G. Fellouris (2017). Asymptotically optimal, sequential, multiple testing procedures with prior information on the number of signals. *Electronic Journal of Statistics 11*(1), 338 – 363.

Song, Y. and G. Fellouris (2019). Sequential multiple testing with generalized error control: An asymptotic optimality theory. *The Annals of Statistics 47*(3), 1776 – 1803.

Tartakovsky, A., I. Nikiforov, and M. Basseville (2014). *Sequential Analysis: Hypothesis Testing and Changepoint Detection* (1st ed.). Chapman & Hall/CRC.

Tsopelakos, A. and G. Fellouris (2023). Sequential anomaly detection under sampling constraints. *IEEE Transactions on Information Theory 69*(12), 8126–8146.

Tsopelakos, A. and G. Fellouris (2025). Sequential anomaly identification under sampling constraints for generalized error metrics. *IEEE Transactions on Information Theory 71*(12), 9753–9783.

Wald, A. (1947). *Sequential Analysis*. New York: John Wiley & Sons.

Wald, A. and J. Wolfowitz (1948). Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics 19*, 326–339.

Xing, Y. (2026). To minimize the expected total sampling cost in sequential testing about a random vector. *Journal of Multivariate Analysis 215*, 105640.

Xing, Y., A. Chaudhuri, and Y. Chen (2025). Signal detection under composite hypotheses with identical distributions for signals and for noises. *arXiv preprint arXiv:2507.21692*.

Xing, Y. and G. Fellouris (2023). Signal recovery with multistage tests and without sparsity constraints. *IEEE Transactions on Information Theory 69*(11), 7220–7245.

Xing, Y. and G. Fellouris (2024). Asymptotically optimal multistage tests for non-iid data. *Statistica Sinica 34*, 2325–2346.

Xing, Y. and G. Fellouris (2025). Asymptotically optimal sequential multiple testing with asynchronous decisions. *Bernoulli 31*(1), 271–294.

Xing, Y. and G. Fellouris (2026). Active sequential signal detection with asynchronous decisions. *arXiv preprint arXiv:2604.04755*.

Zehetmayer, S., P. Bauer, and M. Posch (2005). Two-stage designs for experiments with a large number of hypotheses. *Bioinformatics 21*(19), 3771–3777.

School of Mathematical Sciences, Tongji University, Shanghai, China

E-mail: yimingx4@tongji.edu.cn