ISSN 0005-1179, Automation and Remote Control, 2015, Vol. 76, No. 5, pp. 809–825. © Pleiades Publishing, Ltd., 2015.
Original Russian Text © A.B. Juditsky, A.S. Nemirovski, 2015, published in Avtomatika i Telemekhanika, 2015, No. 5, pp. 100–120.

TOPICAL ISSUE

On Sequential Hypotheses Testing via Convex Optimization

A. B. Juditsky* **and A. S. Nemirovski****

*LJK, Universite Grenoble Alpes, Grenoble, France
**Georgia Institute of Technology, Atlanta, USA
e-mail: anatoli.juditsky@imag.fr, nemirovs@isye.gatech.edu

Received December 1, 2014

Abstract—We propose a new approach to sequential testing which is an adaptive (on-line) extension of the (off-line) framework developed in [1]. It relies upon testing of pairs of hypotheses in the case where each hypothesis states that the vector of parameters underlying the distribution of observations belongs to a convex set. The nearly optimal under appropriate conditions test is yielded by a solution to an efficiently solvable convex optimization problem. The proposed methodology can be seen as a computationally friendly reformulation of the classical sequential testing.

DOI: 10.1134/S0005117915050070

1. INTRODUCTION

Let $\omega_i \in \Omega$, $i = 1, 2, \ldots$, be a sequence of independent and identically distributed (iid) observations with a common density w.r.t. a given measure P on the observation space Ω, known to belong to a given parametric family $\{p_\mu(\cdot) : \mu \in \mathcal{M}\}$. Assume that we are given a (finite) number I of subsets X^i, $1 \leqslant i \leqslant I$, of the parameter space $\mathcal{M}$; these sets define hypotheses H_i on the density $p(\cdot)$ underlying our observation, H_j stating that the parameter of this density belongs to H_j: $p(\cdot) = p_\mu(\cdot)$ with $\mu \in X^i$. We are interested in the problem of multiple testing of composite hypotheses: our goal is, given observations ω_i to decide on the hypotheses $H_1, \ldots, H_I$. When the size K of the observation $\omega^K = (\omega_1, \ldots, \omega_K)$ is fixed, the performance of a test $T^K(\cdot)$—a measurable mapping of Ω^K into $\{1, \ldots, I\}$—can be quantified by the maximal error probability of rejecting the true hypothesis:

$$\mathrm{Risk}(T^K) = \max_{1 \leqslant i \leqslant I} \sup_{\mu \in X^i} \mathrm{Prob}_{\mu,K}\{\omega^K : T^K(\omega^K) \neq i\}, \tag{1}$$

where $\mathrm{Prob}_{\mu,K}$ is the probability w.r.t. the distribution of the observation ω^K corresponding to the "true" parameter $\mu \in \mathcal{M}$. Then the problem of optimal testing can be approached through minimization of the risk over the class of tests. Yet, in many practical applications, especially in those where observations come at a price, such approach may be too conservative. The sequential approach to testing, introduced in the pioneering papers of Barnard [2] and Wald [3, 4] is now a reach area of statistical theory and offers many strong results (see also [5–7] for references or [8] for a recent review). For the problem of interest this approach can be summarized as follows: given an upper bound ϵ of the risk the problem is reexamined sequentially, when new observations are available. The test terminates when either a decision with the risk $\leqslant \epsilon$ is possible, or when maximal allowed observation count K is reached (in which case the test produces no decision). Usually, the sequential test is based on the *Generalized Likelihood* or *Mixture Likelihood* statistics, and its performance is evaluated by its asymptotical, as $\epsilon \to +0$ and $K \to \infty$. The approach we pursuit in this paper is of completely different spirit. It originates from [9] and was applied to the testing problem in [1] in the case where the size of the observation sample is fixed. The main "building block" of this approach is

a construction, based on Convex Programming (and thus computationally efficient) allowing, under appropriate assumptions, to build a provably nearly optimal test for deciding, given observation ω^K, between a pair of composite hypotheses of the sort $H_1 : \mu \in X$ and $H_2 : \mu \in Y$ where X and Y are convex compact subsets of $\mathcal{M}$. This approach is applicable in several important situations, namely where (a) p_μ is Gaussian density on $\mathbb{R}^n$ with expectation μ and fixed covariance matrix, (b) p_μ is the distribution of the Poisson vector in $\mathbb{R}^m$ with independent components with parameters μ_i, and (c) is a distribution of a discrete random variable taking values in $\{1, \ldots, m\}$ with $\mu \in \mathbb{R}_+$ being the vector of probabilities of the corresponding distribution. As a compensation for rather restrictive assumptions on the families of densities p_μ, the approach in question is extremely permissive as far the structure of the sets X and Y is concerned: what we require (apart from compactness and convexity) is the efficient tractability of X and Y. While the analytical expressions for the test characteristics are not available in the proposed approach, rather detailed information about performance guarantees of the resulting decisions can be obtained by efficient situation-oriented computation. In [1] we introduced a calculus of pairwise tests to design nearly optimal testing procedures in the situation where the sets X^i corresponding to different hypotheses H^i are unions of (not too large number of) convex sets with the risk of the test defined as in (1) (in the minimax setting).[1] What follows can be seen as an adaptive version of the testing procedure from [1]—when the "true distribution" of observations corresponds to the value μ of the parameter which is "deeply inside" of some X^i, the correct decision (H_i is accepted) can be taken much faster (using a smaller observation sample) than in the case of μ close to some "wrong" H^i's. On the other hand, it is nothing but a "computationally friendly" application of the classical Sequential Analysis methodology to the problem in question.

The paper is organized as follows. In Section 2 we give a summary of the results of [1] on *hypothesis testing in good observation schemes* and discuss the construction of quasi-optimal "off-line" test. Then in Section 3 we introduce the sequential problem setting and construct a generic test aggregation procedure which allows to reduce multiple testing to pairwise testing. Finally, in Section 4 we consider in detail a particular implementation of the proposed approach and present some very preliminary simulation results. To save the place lengthy proofs are removed from the main body of paper; they may be found in the long version [19] of the manuscript.

2. PRELIMINARIES

What follows is the summary of the approach of [9] as applied to hypotheses testing; for detailed presentation of the constructions and results of this section, same as for the related proofs, see [1].

2.1. Good Observation Schemes

We start with introducing *good observation schemes*, those with which we intend to work. Recall that we are interested to make inferences from a random observation ω taking values in a given *observation space* Ω and obeying probability density w.r.t. a given measure P on Ω; this density is known to belong to a given parametric family $\{p_\mu(\cdot) : \mu \in \mathcal{M}\}$ of probability densities, taken w.r.t. P, on Ω. We intend to work under the following assumptions on our "observation environment":

(1) $\mathcal{M} \subset \mathbb{R}^m$ is a convex set which coincides with its relative interior;

(2) Ω is a Polish (i.e., separable complete metric) space equipped with a Borel σ-additive σ-finite measure P, $\mathrm{supp}(P) = \Omega$, and distributions $P_\mu \in \mathcal{P}$ possess densities $p_\mu(\omega)$ w.r.t. P. We assume that

[1] It should be mentioned that what we call below "simple tests" were used to test composite hypotheses represented by convex sets in the white noise model (a) in [10–12] and in the distribution model (c) in [13–18].

- $p_\mu(\omega)$ is continuous in $\mu \in \mathcal{M}$, $\omega \in \Omega$ and is positive;
- the densities $p_\mu(\cdot)$ are "locally uniformly summable": for every compact set $M \subset \mathcal{M}$, there exists a Borel function $p^M(\cdot)$ on Ω such that $\int_\Omega p^M(\omega)P(d\omega) < \infty$ and $p_\mu(\omega) \leqslant p^M(\omega)$ for all $\mu \in M$, $\omega \in \Omega$;

(3) We are given a finite-dimensional linear space $\mathcal{F}$ of continuous functions on Ω containing constants such that $\ln(p_\mu(\cdot)/p_\nu(\cdot)) \in \mathcal{F}$ whenever $\mu, \nu \in \mathcal{M}$.
Note that the latter assumption implies that distributions P_μ, $\mu \in \mathcal{M}$, belong to an exponential family.

(4) For every $\phi \in \mathcal{F}$, the function $F_\phi(\mu) = \ln\left(\int_\Omega \exp\{\phi(\omega)\}p_\mu(\omega)P(d\omega)\right)$ is well defined and concave in $\mu \in \mathcal{M}$.

In the just described situation, where assumptions (1)–(4) hold, we refer to the collection $\mathcal{O} = ((\Omega, P), \{p_\mu(\cdot) : \mu \in \mathcal{M}\}, \mathcal{F})$ as *good observation scheme* (o.s.).

Basic Examples

Basic examples of good o.s.'s are as follows:

Gaussian o.s. Here $\Omega = \mathbb{R}^m$, P is the measure on Ω, $\mathcal{M} = \mathbb{R}^m$ and $p_\mu(\omega) = \mathcal{N}(\mu, I_m)$ is the density of Gaussian random vector with the unit covariance matrix[2] and expectation μ. The space $\mathcal{F}$ is comprised of all affine functions on $\Omega = \mathbb{R}^m$, and because

$$\ln\left(\int_{\mathbb{R}^m} e^{a^{\mathrm{T}}\omega+b}p_\mu(\omega)d\omega\right) = b + a^{\mathrm{T}}\mu + \frac{1}{2}a^{\mathrm{T}}a,$$

Gaussian o.s. is good.

Poisson o.s. Here $\Omega = \mathbb{Z}^m_+$ is the discrete set of m-dimensional vectors with nonnegative integer entries, P is the counting measure on Ω, $\mathcal{M} = \mathbb{R}^m_{++}$ is the set of positive m-dimensional vectors, and $p_\mu(\cdot)$, $\mu = [\mu_1; \ldots; \mu_m] > 0$, is the probability distribution of a random vector $\omega = [\omega_1; \ldots; \omega_m]$, where ω_i is Poisson random variable with parameter μ_i, and with independent of each other $\omega_1, \ldots, \omega_m$. The space $\mathcal{F}$ is comprised of all affine functions on Ω. Note that

$$\ln\left(\sum_{\omega\in\mathbb{Z}^m_+} \exp(a^{\mathrm{T}}\omega + b)p_\mu(\omega)\right) = \sum_{i=1}^m (e^{a_i} - 1)\mu_i + b$$

is concave in μ, and thus Poisson o.s. is good.

Discrete o.s. Here $\Omega = \{1, \ldots, m\}$ is a finite set, P is a counting measure on Ω, and $\mathcal{M}$ is comprised of all non-vanishing probability densities taken w.r.t. P, that is, the parameter space $\mathcal{M}$ is comprised of all m-dimensional vectors $\mu = [\mu_1; \ldots; \mu_m] > 0$ with entries summing up to 1, and the random variable ω distributed according to $p_\mu(\cdot)$ takes value $\omega \in \Omega$ with probability μ_ω. The space $\mathcal{F}$ is comprised of all real-valued functions on $\Omega = \{1, \ldots, m\}$. Since for $\phi \in \mathbb{R}^m$,

$$\ln\left(\sum_{\omega\in\Omega} e^{\phi(\omega)}p_\mu(\omega)\right) = \ln\left(\sum_{\omega=1}^m e^{\phi_\omega}\mu_\omega\right)$$

is concave in $\mu \in \mathcal{M}$, the Discrete o.s. is good.

We have seen that Gaussian, Poisson, and Discrete o.s.'s are good. More examples of good o.s.'s can be obtained by taking *direct products*.

Direct products of good o.s.'s. Let $\mathcal{O}_t = ((\Omega_t, P_t), \{p_{\mu_t,t}(\cdot) : \mu_t \in \mathcal{M}_t\}, \mathcal{F}_t)$, $1 \leqslant t \leqslant K$, be good o.s.'s. We can associate with this collection a new o.s.—their *direct product*, which, informally,

[2] Of course, we could replace the unit covariance with any other positive definite covariance matrix, common for all distributions from the family in question.

describes the situation where our observation is $\omega^K = (\omega_1, \ldots, \omega_K)$, with ω_t drawn, independently across t, from the o.s.'s $\mathcal{O}_t$. Formally, the direct product is the o.s.

$$\mathcal{O}^K = \left\{(\Omega^K, P^K) = \left(\prod_{t=1}^K \Omega_t, P^K = P_1 \times \ldots \times P_K\right),\right.$$
$$\left\{p_{\mu^K}(\omega_1, \ldots, \omega_K) = p_{\mu_1,1}(\omega_1)\ldots p_{\mu_K,K}(\omega_K) : \mu^K = [\mu_1; \ldots; \mu_K] \in \mathcal{M}^K = \mathcal{M}_1 \times \ldots \times \mathcal{M}_K\right\},$$
$$\left.\mathcal{F}^K = \left\{f(\omega_1, \ldots, \omega_K) = \sum_{t=1}^K f_t(\omega_t) : f_t \in \mathcal{F}_t, 1 \leqslant t \leqslant K\right\}\right\},$$

and this o.s. turns to be good provided all factor $\mathcal{O}_t$ are so. Note that with this definition, the parameter μ^K underlying the distribution of observation ω^K is the collection of parameters μ_t, $t = 1, \ldots, K$, underlying distributions of the components $\omega_1, \ldots, \omega_K$ of ω^K.

Now consider the special case of this construction where all factors $\mathcal{O}_t$ are identical to each other:

$$\mathcal{O}_t = ((\Omega, P), \{p_\mu(\cdot) : \mu \in \mathcal{M}\}, \mathcal{F}),\ 1 \leqslant t \leqslant K.$$

In this case we can "shrink" the direct product $\mathcal{O}^K$ of our K identical o.s.'s to the o.s. $\mathcal{O}^{(K)}$ by passing to observations $\omega^K = (\omega_1, \ldots, \omega_K)$ with $\omega_1, \ldots, \omega_K$ drawn, independently of each other, from *the same* density $p_\mu(\cdot)$, rather than being drawn from their own densities $p_{\mu_t,t}(\cdot)$. The formal description of $\mathcal{O}^{(K)}$ is as follows:

- for $\mathcal{O}^{(K)}$, the observation space $\Omega^K = \Omega \times \ldots \times \Omega$ and the reference measure $P^K = P \times \ldots \times P$ are exactly the same as for $\mathcal{O}^K$;
- the family of probability densities $\{p(\cdot)\}$ for $\mathcal{O}^{(K)}$ is $\{p_\mu^{(K)}(\omega_1, \ldots, \omega_K) = \prod_{t=1}^K p_\mu(\omega_t) : \mu \in \mathcal{M}\}$;
- the family $\mathcal{F}$ for $\mathcal{O}^{(K)}$ is $\mathcal{F}^{(K)} = \{f^{(K)}(\omega_1, \ldots, \omega_K) = f(\omega_1) + \ldots + f(\omega_K) : f \in \mathcal{F}\}$.

Note that $\mathcal{O}^{(K)}$ is a good o.s., provided $\mathcal{O}$ is so. We shall refer to the just defined $\mathcal{O}^{(K)}$ as to *stationary K-repeated observations* associated with $\mathcal{O}$.

2.2. Pairwise Hypothesis Testing

Detectors and their risks. Let Ω be a Polish space and $\mathcal{X}_1, \mathcal{X}_2$ be two nonempty sets of Borel probability distributions on Ω. Given a *detector*—a real-valued Borel function ϕ on Ω, we define the *risk* $\epsilon(\phi|\mathcal{X}_1, \mathcal{X}_2)$ of this detector w.r.t. $(\mathcal{X}_1, \mathcal{X}_2)$ as the smallest ϵ such that

$$\begin{aligned}\int_\Omega \exp\{-\phi(\omega)\}p(d\omega) &\leqslant \epsilon\ \forall p \in \mathcal{X}_1,\\ \int_\Omega \exp\{\phi(\omega)\}p(d\omega) &\leqslant \epsilon\ \forall p \in \mathcal{X}_2,\end{aligned} \tag{2}$$

The sets $\mathcal{X}_1, \mathcal{X}_2$ give rise to two hypotheses, H_1, H_2, on the distribution of a random observation $\omega \in \Omega$, with H_χ stating that the distribution $p(\cdot)$ of this observation belongs to $\mathcal{X}_\chi$, $\chi = 1, 2$, while a detector ϕ gives rise to the test T_ϕ which, given an observation ω, accepts H_1 (and rejects H_2) when $\phi(\omega) \geqslant 0$, and rejects H_1 (and accepts H_2) otherwise. We define *the risk of a test T* deciding between H_1 and H_2 as a maximal $p(\cdot)$-probability of the test rejecting the hypothesis H_χ, $\chi = 1, 2$, when it is true:

$$\text{Risk} = \max\left[\sup_{p(\cdot)\in\mathcal{X}_1} p(\{\omega : T(\omega) = -1\}), \sup_{p(\cdot)\in\mathcal{X}_2} p(\{\omega : T(\omega) = 1\})\right],$$

where $T(\omega) = 1$ when the test T, as applied to observation ω, accepts H_1, and $T(\omega) = -1$ otherwise. Clearly, the risk of the test T_ϕ associated with detector ϕ is $\leqslant \epsilon := \epsilon(\phi|\mathcal{X}_1, \mathcal{X}_2)$. Indeed, denoting

by $p(\cdot)$ the distribution of the observation, when H_1 is true, the test rejects this hypothesis when $\phi(\omega) < 0$, and $p(\cdot)$-probability of this event, by the first inequality in (2) and due to the definition of $\epsilon := \epsilon(\phi|\mathcal{X}_1, \mathcal{X}_2)$, is at most ϵ; when H_2 is true, the test rejects H_2 and accepts H_1 when $\phi(\omega) \geqslant 0$, and $p(\cdot)$-probability of the latter event is $\leqslant \epsilon$ due to the second inequality in (2).

Note that if we have at our disposal a test, say $\overline{T}$, which decides between H_1 and H_2 with the risk bounded with $\bar{\epsilon} \in (0, 1/2)$, we can associate with it the detector

$$\bar{\phi}(\omega) = \frac{1}{2} \ln \left(\frac{1 - \bar{\epsilon}}{\bar{\epsilon}} \right) \overline{T}(\omega)$$

One can easily see that the risk of $\bar{\phi}(\cdot)$ satisfies the bounds of (2) with

$$\epsilon = 2\sqrt{\bar{\epsilon}(1 - \bar{\epsilon})}.$$

As we will see in an instant, tests associated with detectors satisfying (2) allow for a simple calculus—one can "propagate" the tests properties to the case of repeated observations and multiple testing. Our first goal is to describe a systematic construction of detectors satisfying (2) in the situation where the underlying o.s. is good.

Near-optimal tests. As far as testing pairs of hypotheses is concerned, the main result of [1]—and the starting point of our developments in this paper—is as follows:

Theorem [1, Theorem 2.1]. *Let $\mathcal{O} = ((\Omega, P), \{p_\mu(\cdot) : \mu \in \mathcal{M}\}, \mathcal{F})$ be a good o.s., and let X_1, X_2 be two nonempty convex compact subsets of $\mathcal{M}$. Consider the optimization problem*

$$\mathrm{Opt} = \mathrm{Opt}(X_1, X_2, \mathcal{O}) := \max_{\mu \in X_1, \nu \in X_2} \left\{ \psi(\mu, \nu) := \ln \left(\int_\Omega \sqrt{p_\mu(\omega) p_\nu(\omega)} P(d\omega) \right) \right\}. \tag{3}$$

The function $\psi(\mu, \nu) : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ is concave and continuous, so that (3) *is a convex optimization problem. This problem is solvable, and every optimal solution (μ_*, ν_*) to this problem gives rise to the detector*

$$\phi(\omega) = \frac{1}{2} \ln(p_{\mu_*}(\omega)/p_{\nu_*}(\omega)) : \Omega \to \mathbb{R} \tag{4}$$

with the following properties:

(i) [risk bound] *Denoting by $\mathcal{X}_\chi$ the set of all probability distributions on Ω with densities, w.r.t. P, of the form $p_\mu(\cdot)$ with $\mu \in X_\chi$, $\chi = 1, 2$, the risk of the detector ϕ w.r.t. $(\mathcal{X}_1, \mathcal{X}_2)$ is*

$$\epsilon(\phi|\mathcal{O}, \mathcal{X}_1, \mathcal{X}_2) := \epsilon_\star(\mathcal{O}, X_1, X_2) = \exp\{\mathrm{Opt}\};$$

consequently, the risk of the test T_ϕ induced by the detector ϕ, when deciding on the hypotheses H_1 and H_2 associated with $\mathcal{X}_1$ and $\mathcal{X}_2$, as at most $\exp\{\mathrm{Opt}\}$.

(ii) [near-optimality] *Assume that for some $\epsilon \in (0, 1/4)$, "in the nature" there exists a test T (a deterministic Borel function $T(\omega)$ of $\omega \in \Omega$) taking values 1 ("H_1 accepted, H_2 rejected") and -1 ("H_2 accepted, H_1 rejected") with risk in deciding on H_1, H_2 based on an observation ω at most ϵ. Then*

$$\epsilon_*(X_1, X_2) \leqslant 2\sqrt{\epsilon}.$$

We interpret (ii) as a statement of near-optimality of the test T defined in (i)—whenever the hypotheses H_1, H_2 associated with $\mathcal{X}_1$, $\mathcal{X}_2$ can be decided upon with small risk, the risk of our test T_ϕ also is small. Note that (ii) also suggests that (maximal) degradation of the test performance when passing from the optimal test to the near-optimal one associated with detector (4) may be significant in the setting of theorem, when the decision is taken on the basis of one observation. When repeated observations are available, the sub-optimality of the proposed tests is expressed by a moderate absolute factor, when measured in terms of the length of the observation sample necessary to attain the desired testing accuracy.

Corollary 1. *Let $\mathcal{O} = ((\Omega, P), \{p_\mu(\cdot) : \mu \in \mathcal{M}\}, \mathcal{F})$, X_1, X_2 be as in theorem, and K be a positive integer. Assume we have at our disposal stationary K-repeated observation ω^K stemming from $\mathcal{O}$, and let H_χ, $\chi = 1, 2$, be the hypotheses on the density $p_\mu(\omega_1, \ldots, \omega_K) = \prod_{t=1}^K p_\mu(\omega_t)$ of ω^K stating that $\mu \in X_\chi$. Then*

(i) *The optimal solution (μ_*, ν_*) to the problem (3) associated with $\mathcal{O}$, X_1, X_2, is optimal for the same problem associated with $\mathcal{O}^{(K)}, X_1, X_2$, and*

$$\epsilon_\star(\mathcal{O}^{(K)}, X_1, X_2) = [\epsilon_\star(\mathcal{O}, X_1, X_2)]^K;$$

moreover, the detectors ϕ and $\phi^{(K)}$ associated with this optimal solution by theorem as applied to $(\mathcal{O}, X_1, X_2)$ and $(\mathcal{O}^K, X_1, X_2)$, respectively, are linked by the relation

$$\phi^{(K)}(\omega_1, \ldots, \omega_K) = \sum_{t=1}^K \phi(\omega_t).$$

(ii) *Assume that for some $\epsilon \in (0, 1/4)$ and some positive integer $\bar{K}$, there exists a test which decides on the hypotheses H_1, H_2 via stationary $\bar{K}$-repeated observation $\omega^{\bar{K}}$ with risk $\leqslant \epsilon$. Setting*

$$K = \mathrm{Ceil}\left(\frac{2\bar{K}}{1 - \frac{2\ln(2)}{\ln(1/\epsilon)}}\right),$$

we ensure that the risk of the test based on detector $\phi^{(K)}(\cdot)$ when deciding on the hypotheses H_χ, $\chi = 1, 2$, is $\leqslant \epsilon$ as well. Note that $K/\bar{K} \to 2$ as $\epsilon \to +0$.

2.3. Testing Multiple Hypotheses

The completing "building block" from [1] we need is a simple technique for passing from pairwise tests to tests deciding on $N \geqslant 2$ hypotheses.

The situation we are interested in is as follows. We are given a Polish observation space Ω along $N \geqslant 2$ nonempty families $\mathcal{X}_j$, $1 \leqslant j \leqslant N$, of Borel probability distributions on Ω, and have at our disposal *pairwise detectors*—real-valued Borel functions $\phi_{ij}(\omega) : \Omega \to \mathbb{R}$ and *risk bounds* $\epsilon_{ij} \in (0, 1]$, $1 \leqslant i, j \leqslant N$, such that

$$\begin{gathered} \phi_{ij}(\cdot) = -\phi_{ji}(\cdot), \quad \epsilon_{ji} = \epsilon_{ij}, \quad 1 \leqslant i, j \leqslant N, \\ \int_\Omega \exp\{-\phi_{ij}(\omega)\} p(d\omega) \leqslant \epsilon_{ij} \ \forall p(\cdot) \in \mathcal{X}_i, \quad 1 \leqslant i, j \leqslant N. \end{gathered} \tag{5}$$

Our goal is, given a positive integer K, to decide via stationary K-repeated observation $\omega^K = (\omega_1, \ldots, \omega_K)$, with $\omega_1, \ldots, \omega_K$ drawn, independently of each other, from a distribution $p(\cdot) \in \bigcup_{j=1}^M \mathcal{X}_j$, between hypotheses H_j, $1 \leqslant j \leqslant N$, with H_j stating that $p(\cdot) \in \mathcal{X}_j$.

Note that the present setting is a straightforward extension of the situation considered in Section 2.2. In particular, given what was called in this section "a detector ϕ with (X_1, X_2)-risk $\leqslant \epsilon$," and setting $\phi_{1,2} = \phi$, $\phi_{2,1} = -\phi$, $\phi_{1,1} \equiv \phi_{2,2} \equiv 0$, $\epsilon_{1,2} = \epsilon_{2,1} = \epsilon$, $\epsilon_{1,1} = \epsilon_{2,2} = 1$, we meet the requirements (5) corresponding to the case $N = 2$. Vice versa, given detectors and risks satisfying (5) for the case $N = 2$ and setting $\phi = \phi_{1,2}$, $\epsilon = \epsilon_{1,2}$, we get a detector ϕ with (X_1, X_2)-risk $\leqslant \epsilon$.

We are about to "aggregate" detectors ϕ_{ij} into a testing procedure deciding on the hypotheses $H_1, \ldots, H_N$ via stationary K-repeated observations ω^K. It makes sense to consider a slightly more general problem, specifically, as follows: assume that on the top of the setup data $\mathcal{X}_j$, $\phi_{ij}(\cdot)$, ϵ_{ij}, we are given an $N \times N$ symmetric (proximity) matrix $\mathcal{C}$ with zero-one entries and zero diagonal.

We interpret the relation $\mathcal{C}_{ij} = 0$ as *closeness* of hypotheses H_i and H_j,[3] and we refer to indices i, j (and hypotheses H_i, H_j) such that $\mathcal{C}_{ij} = 0$ as *$\mathcal{C}$-close.* We expect our testing procedure not to reject the true hypothesis, while rejecting all hypotheses which are not *not* $\mathcal{C}$-close to it. On the other hand, we do not care about distinguishing the true hypothesis from $\mathcal{C}$-close alternatives.

The construction. Given $\mathcal{X}_j$, $\phi_{ij}(\cdot)$, ϵ_{ij} satisfying (5) along with positive integer K, let us set

$$\phi_{ij}^K(\omega^K) = \sum_{t=1}^K \phi_{ij}(\omega_t), \quad \epsilon_{ij}^{(K)} = \epsilon_{ij}^K,$$

and let $\mathcal{X}_j^K$, $j = 1, \ldots, N$ be a family of probability distributions of $\omega^K = (\omega_1, \ldots, \omega_K)$ where ω_t are i.i.d. with distribution $p(\cdot) \in \mathcal{X}_j$. Clearly, (5) implies that

$$\phi_{ij}^K(\cdot) = -\phi_{ji}^K(\cdot),\ \epsilon_{ji}^{(K)} = \epsilon_{ij}^{(K)}, \quad 1 \leqslant i, j \leqslant N$$
$$\int_{\Omega^K} \exp\{-\phi_{ij}^K(\omega^K)\} p(d\omega^K) \leqslant \epsilon_{ij}^{(K)}\ \forall p(\cdot) \in \mathcal{X}_i^K, \quad 1 \leqslant i, j \leqslant N.$$

Now, let $\alpha = [\alpha_{ij}]$ be a skew-symmetric $N \times N$ matrix, and let

$$\overline{\phi}_{ij}(\omega^K) = \phi_{ij}^K(\omega^K) - \alpha_{ij}.$$

We associate with $\overline{\phi}$ the test $\mathcal{T}_K$ which, given observation ω^K, builds the $N \times N$ matrix with the entries $\overline{\phi}_{ij}(\omega^K)$ and accepts all hypotheses H_i which satisfy the condition

$$\overline{\phi}_{ij}(\omega^K) > 0\ \ \forall (j : C_{ij} = 1),$$

and rejects all remaining hypotheses. Note that $\mathcal{T}_K$ can accept no hypotheses at all, or can accept more than one hypothesis.

The properties of $\mathcal{T}_K$ are summarized in the following simple statement (see [1, Section 2.3.1]):

Proposition 1. *Let*

$$\varepsilon = \max_{1 \leqslant i \leqslant N} \sum_{j : \mathcal{C}_{ij} = 1} \epsilon_{ij}^K \exp\{-\alpha_{ij}\}. \tag{6}$$

Let $\omega^K = (\omega_1, \ldots, \omega_K)$ be sampled independently from the distribution $p_*(\cdot) \in \mathcal{X}_{i_*}$, for some $i^* \in \{1, \ldots, N\}$. Then

(i) *the $p_*(\cdot)$-probability of $\mathcal{T}_K$ rejecting the true hypothesis H_{i_*} is $\leqslant \varepsilon$;*

(ii) *the $p_*(\cdot)$-probability of the event "among the hypotheses accepted by $\mathcal{T}_K$, there are hypotheses which are not $\mathcal{C}$-close to the true hypothesis H_{i_*}," is $\leqslant \epsilon$.*

The risk bound ε, as given by (6), depends on the "shifts" $\alpha_{ij} = -\alpha_{ji}$, and we would like to use the shifts resulting in as small as possible value of ε in (1). It is shown in [1, Section 3] that the corresponding shifts solve the convex optimization problem

$$\mathrm{Opt} = \min_{\alpha \in \mathbb{R}^{N \times N}} \left\{ f(\alpha) = \max_{1 \leqslant i \leqslant N} \sum_{j : \mathcal{C}_{ij} = 1} \epsilon_{ij}^K \exp\{\alpha_{ij}\} : \alpha = -\alpha^{\mathrm{T}} \right\}, \tag{7}$$

the optimal ε in (6) being exactly Opt. Moreover, from [1, Proposition 3.3] it immediately follows (see below), that Opt is nothing but the spectral norm $\|D\|_{2,2}$ of the entry-wise nonnegative symmetric matrix

$$D = [d_{ij} = \epsilon_{ij}^K \mathcal{C}_{ij}]_{1 \leqslant i, j \leqslant N}.$$

[3] In the general case considered in [1], the closeness (i.e., the zero-one matrix $\mathcal{C}$ with zero diagonal) not necessary is symmetric; here we restrict ourselves with symmetric closeness only.

Moreover, it is immediately seen (cf. [1, Section 3.2]) that *if the Perron–Frobenius eigenvector g of the entrywise nonnegative symmetric matrix D is positive, an optimal solution to* (7) *is given by*

$$\alpha_{ij} = \ln(g_j) - \ln(g_i). \tag{8}$$

In the general case the Perron–Frobenius eigenvector of D can have zero entries, and, moreover, (7) may happen not to have optimal solutions at all; we, however, can easily find ϵ-optimal, with a whatever small $\epsilon > 0$, solutions to the problem by utilizing in (8) in the role of g the Perron-Frobenius eigenvector of a matrix D' with the entries $d'_{ij} > d_{ij}$ close to d_{ij}, specifically, such that $\|D'\|_{2,2} \leqslant \|D\|_{2,2} + \epsilon$.

In fact, Proposition 3.3 from [1] states that if $E = [e_{ij}]$ is a symmetric $N \times N$ matrix with zero diagonal and *positive* off-diagonal entries, then the optimal value in the optimization problem

$$\mathrm{Opt}_E = \min_{\alpha} \left\{ \max_{1 \leqslant i \leqslant N} \sum_{j=1}^{N} e_{ij} \exp\{\alpha_{ij}\} : \alpha = -\alpha^{\mathrm{T}} \right\}$$

is equal to the spectral norm $\|E\|_{2,2}$. Let now $e_{ij} = d_{ij} (= \epsilon_{ij}^K \mathcal{C}_{ij})$ for $1 \leqslant i, j \leqslant N$ such that $\mathcal{C}_{ij} = 1$, and let $e_{ij} = \epsilon > 0$ for $i \neq j$ such that $\mathcal{C}_{ij} = 0$, so that all off-diagonal entries of E are positive. Note that when $\epsilon \downarrow 0$, both the spectral norms of $\|E\|_{2,2}$ and $\|D\|_{2,2}$, and the optimal values Opt and Opt_E become arbitrarily close to each other. Since for the "perturbed" matrix E the spectral norm $\|E\|_{2,2}$ coincides with the optimal value Opt_E, the same holds true for the "unperturbed" matrix D.

3. SEQUENTIAL HYPOTHESIS TESTING

3.1. Problem Setting

Let us consider the situation of Section 2.3. Specifically, assume that we are given $N \geqslant 2$ nonempty families $\mathcal{X}_j$ of Borel probability distributions on a Polish observation space Ω. Let, further,

$$\mathcal{J} := \{1, 2, \ldots, N\} = \bigcup_{i=1}^{I} \mathcal{J}_i$$

be a partition of the set of indices of X_j's into $I \geqslant 2$ non-overlapping nonempty groups $\mathcal{J}_1, \ldots, \mathcal{J}_I$. We associate with the sets $\mathcal{X}^i = \bigcup_{j \in \mathcal{J}_i} \mathcal{X}_j$ the hypotheses H_i, stating that the elements $\omega_1, \ldots, \omega_K$ of the K-repeated stationary observation sample ω^K are drawn independently from the common distribution $p \in \mathcal{X}^i$, and our goal is to decide from observation ω^K on the hypotheses $H_1, \ldots, H_I$.

It is convenient to think about the (values of the) indices $i = 1, \ldots, I$ of the sets $\mathcal{X}^i$ as of the *colors* of these sets.[4]

We also assume that we have at our disposal pairwise detectors—Borel functions $\phi_{ij} : \Omega \to \mathbb{R}$—for the sets $\mathcal{X}_j$, $j = 1, \ldots, N$, which satisfy relations (5). From now on, we make the following assumption:

A: *When the indices j and $j' \in \mathcal{J}$ are of different colors (in other words, do not belong to the same group $\mathcal{J}_i$), the risk $\epsilon_{jj'} (= \epsilon_{j'j})$ of the detector $\phi_{jj'}$ (same as the detector $\phi_{j'j}$) satisfies $\epsilon_{jj'} < 1$.*

[4] Let us agree that these colors are inherited by different entities associated with $\mathcal{X}^i$'s such as index groups $\mathcal{J}_i$, indices $j \in \mathcal{J}_i$, corresponding sets $\mathcal{X}_j$ and distributions $p \in \mathcal{X}^i$.

Note that assumption **A** implies that for j and j' of different colors the sets of distributions $\mathcal{X}_j$ and $\mathcal{X}_{j'}$ are at positive Hellinger distance from each other:

$$\inf_{p(\cdot)\in\mathcal{X}_j, q(\cdot)\in\mathcal{X}_{j'}} \int_\Omega \left(\sqrt{p(d\omega)} - \sqrt{q(d\omega)}\right)^2 > 0.$$

This implies that our color assignments are unambiguous, and our goal may be reformulated as that of identifying the color of the distribution underlying the observations.

From the results of Section 2.3 it easily follows that if assumption **A** holds then for any given $\epsilon > 0$, we can decide on the hypotheses $H_1, \ldots, H_I$ with risk $\leqslant \epsilon$ (meaning that the probability to reject the true hypothesis, same as the probability to accept a wrong one, is $\leqslant \epsilon$), provided that the number K of observations is large enough. This being said, the "large enough" K could be indeed quite large for the pairs $j \in \mathcal{J}_i$, $j' \in \mathcal{J}_{i'}$ with $i \neq i'$, with small value of $\epsilon_{jj'}$. As a tradeoff, we can switch from decision rules based on K observations to *sequential* decision rules, for which the decision is made on the basis of on-line adjustable number of observations. We can expect that if we are lucky and the distribution p_* underlying our observation is "deeply inside" of some $\mathcal{X}^{i_*}$ and thus is "far" from all $\mathcal{X}^i$, $i \neq i_*$, the true hypothesis H_{i_*} will be accepted much sooner than in the case when p_* is close to some of "wrong" $\mathcal{X}^i$'s. Our objective now is to build sequential tests utilizing the results of Section 2.

3.2. Sequential Test: Construction

The setup for our "generic" sequential test is given by

(1) required *risk* $\epsilon \in (0,1)$;
(2) positive integer S—*number of stages*, along with the following entities, defined for $1 \leqslant s \leqslant S$ and forming *sth component* of the setup:
   (a) positive reals ϵ_s, $1 \leqslant s \leqslant S$, such that $2\sum_{s=1}^S \epsilon_s = \epsilon$;
   (b) representations $\mathcal{X}_j = \bigcup_{\iota=1}^{\iota_{js}} \mathcal{X}_{j\iota s}$, $j \in \mathcal{J}$, where $\mathcal{X}_{j\iota s}$ are nonempty closed convex subsets of $\mathcal{X}_j$;
   (c) *tolerances* $\delta_s \in (0,1)$.

To avoid messy notation, we enumerate, for every $s \leqslant S$, the sets $\mathcal{X}_{j\iota s}$, $1 \leqslant j \leqslant N$, $1 \leqslant \iota \leqslant \iota_{js}$, and call the resulting sets $\mathcal{Z}_{1s}, \ldots, \mathcal{Z}_{L_s s}$. Thus, $\mathcal{Z}_{qs}$ is one of the sets $\mathcal{X}_{j\iota s}$, and we assign to $\mathcal{Z}_{qs}$ and the index q the same color as that of j.

Detectors. We suppose for every $s \leqslant S$ and every pair (q, q'), $1 \leqslant q < q' \leqslant L_s$, we are given pairwise detectors $\phi_{qq',s}$ and reals $\epsilon_{qq',s} \in (0,1]$ associated with $\mathcal{Z}_{qs}$ and $\mathcal{Z}_{q's}$ and satisfying the relations

$$\phi_{qq',s}(\cdot) = -\phi_{q'q,s}(\cdot), \quad \epsilon_{qq',s} = \epsilon_{q'q,s}, \quad 1 \leqslant q, q' \leqslant L_s$$
$$\int_\Omega \exp\{-\phi_{qq',s}(\omega)\} p(d\omega) \leqslant \epsilon_{qq',s} \ \forall p(\cdot) \in \mathcal{Z}_{qs}, \quad 1 \leqslant q, q' \leqslant L_s.$$

***s*-Closeness.** Let H_{qs} be the hypotheses on the distribution $p(\cdot)$ of an observation stating $p(\cdot) \in \mathcal{Z}_{qs}$; by convention, H_{qs} is of the same color as $\mathcal{Z}_{qs}$. Let us say that hypothesis $H_{q's}$ is *s-close* to hypothesis H_{qs} (same as q' is s-close to q), if either H_{qs} and $H_{q's}$ are of the same color, or $\epsilon_{qq',s} > \delta_s$ if they are of different colors. Let $\mathcal{C} = \mathcal{C}^s$ be the $L_s \times L_s$ matrix with (q,q')-entry equal to 0 if and only if q is s-close to q', and equal to 1 otherwise. This matrix clearly meets the requirements imposed on $\mathcal{C}$ in Section 2.3: it is a symmetric matrix (recall that $\epsilon_{qq',s} = \epsilon_{q'q,s}$) with 0/1 entries and zero diagonal.

Tests T_s. Now let us apply to the collection of hypotheses $\{H_{qs} : 1 \leqslant q \leqslant L_s\}$, detectors $\{\phi_{qq',s}(\cdot) : 1 \leqslant q, q' \leqslant L_s\}$ and the just defined matrix $\mathcal{C}^s$ the construction from Section 2.3, assuming that

when deciding upon hypotheses $H_{1s},\dots,H_{L_ss}$, we have at our disposal k-repeated observation ω^k, with a given k. Specifically, consider the optimization problem

$$\mathrm{Opt}(k,s)=\min_{\alpha\in\mathbb{R}^{L_s\times L_s}}\left\{f_s(\alpha):=\max_q\sum_{q':\mathcal{C}^s_{qq'}=1}\epsilon^k_{qq',s}\exp\{\alpha_{qq'}\}:\alpha=-\alpha^{\mathrm{T}}\right\};\tag{9}$$

as was shown in Section 2.3, the value $\mathrm{Opt}(k,s)$ is nothing but the spectral norm of the entry-wise nonnegative symmetric matrix with the entries $\epsilon^k_{qq',s}\mathcal{C}^s_{qq'}$.

Since $\epsilon_{qq',s}\leqslant\delta_s\in(0,1)$ when $\mathcal{C}^s_{qq'}=1$, $\mathrm{Opt}(k,s)$ goes to 0 as $k\to\infty$, so that the smallest $k=k(s)$ such that $\mathrm{Opt}(k,s)<\epsilon_s$ is well defined. And since $\mathrm{Opt}(k(s),s)<\epsilon_s$, problem (9) with $k=k(s)$, whether solvable or not, admits a feasible solution $\bar\alpha^{(s)}$ such that $f_s(\bar\alpha^{(s)})\leqslant\epsilon_s$. Applying to detectors $\overline{\phi}_{qq',s}(\cdot)=\phi_{qq',s}(\cdot)-\bar\alpha^{(s)}_{qq'}$ and to $\mathcal{C}=\mathcal{C}^s$ the construction from Section 2.3, we get a test T_s deciding on the hypotheses H_{qs}, $1\leqslant q\leqslant L_s$, via $k(s)$-repeated observation $\omega^{k(s)}$, with properties as follows:

Let $\omega_1,\dots,\omega_{k(s)}$ be drawn, independently of each other, from common distribution $p_*(\cdot)$ obeying the hypothesis H_{q_*s} for some $1\leqslant q_*\leqslant L_s$. Then

1. *the p_*-probability for T_s not to accept H_{q_*s} is at most ϵ_s;*
2. *the p_*-probability of the event "among the accepted hypotheses, there is a hypothesis H_{qs} with q not s-close to q_*" is $\leqslant\epsilon_s$.*

***S*th setup component.** We assume that *when $s=S$, the partition of $\mathcal{X}_j$ is trivial: $\iota_{jS}=1$ and $\mathcal{X}_{j1S}=\mathcal{X}_j$ for all $j\leqslant N$.* Furthermore, we define $K=k(S)$ such that $\delta_S\geqslant\epsilon_{qq',S}$ *whenever q, q' are of different colors*, implying that *q and q' are S-close if and only if q and q' are of the same color.*[5] For the reasons which will become clear in a moment, we are not interested in those components of our setup for which $k(s)>K$; if components with this property were present in our original setup, we can just eliminate them, reducing S accordingly. Finally, we can reorder the components of our setup to make $k(s)$ nondecreasing in s. Thus, from now on we assume that

$$k(1)\leqslant k(2)\leqslant\ldots\leqslant k(S)=:K.$$

Sequential test $\mathcal{T}$ corresponding to the outlined setup when applied to the observation ω^K works by stages $s=1,2,\dots,S$: at stage s we apply test T_s to the initial fragment $\omega^{k(s)}$ of ω^K. If the outcome of the latter test is acceptance of a nonempty set of hypotheses H_{qs} *and all these hypotheses are of the same color i*, test $\mathcal{T}$ accepts the hypothesis $\mathcal{H}_i$ and terminates, otherwise it proceeds to stage $s+1$ (when $s<S$) or terminates without accepting any hypothesis ($s=S$).

Note that by construction $\mathcal{T}$ never accepts more than one of the hypotheses $H_1,\dots,H_I$.

3.3. Sequential Test: Analysis

For a distribution $p\in\mathcal{X}=\bigcup_{j=1}^N\mathcal{X}_j$, let $s[p]\in\{1,2,\dots,S\}$ be defined as follows: for every s, p belongs to (perhaps, several) of the sets $\mathcal{Z}_{qs}$, $1\leqslant q\leqslant L_s$. Further, let $Q_s[p]$ be the set of all q's such that $p\in\mathcal{Z}_{qs}$. Note that the color of every $q\in Q_s[p]$ (recall that the sets $\mathcal{Z}_{qs}$ and the corresponding values of q have already been assigned colors) is the same as the color of p. Now, given $p\in\mathcal{X}$, for some $s\leqslant S$ it may happen that

$$\exists q\in Q_s[p]:\ \text{all } q'\ s\text{-close to } q \text{ are of the same color as } q.\tag{10}$$

[5] We can do so because by assumption **A** *all quantities $\epsilon_{qq',s}$, $1\leqslant q,q'\leqslant L_S=N$, with q, q' of different colors are less than 1.*

In particular, the latter condition definitely takes place when $s = S$, since, as we have already seen, q and q' are S-close if and only if q and q' are of the same color. Hence for $s = S$ the conclusion in (10) is satisfied for all $q \in Q_S[p]$. Now let $s[p]$ be the smallest s such that (10) takes place, and let the corresponding q be denoted by $q[p]$. Thus for all $p \in \mathcal{X}$,

$$s[p] \in \{1, 2, \ldots, S\},\ q[p] \in \{1, \ldots, L_{s[p]}\},\ p \in \mathcal{Z}_{q[p]s[p]};$$

whenever $q' \in \{1, \ldots, L_{s[p]}\}$ is $s[p]$-close to $q[p]$, q' and $q[p]$ are of the same color.

The main result of this section is as follows.

Proposition 2. *Let $\omega_1, \ldots, \omega_K$ be drawn, independently of each other, from a distribution $p_* \in \mathcal{X}(= \bigcup_{j=1}^N \mathcal{X}_j)$, so that $p_* \in \mathcal{X}_{j_*}$ for some $j_* \leqslant N$, and let i_* be such that $j_* \in \mathcal{J}_{i_*}$ (i.e., i_* is the color of p_*). Then p_*-probability of the event*

$$\mathcal{E} = \left\{\omega^K : \begin{array}{l} \mathcal{T}, \text{ as applied to } \omega^K, \text{ terminates not later than at the stage } s[p_*] \\ \text{and accepts upon termintation the true hypothesis } \mathcal{H}_{i_*} \end{array}\right\}$$

is at least $1 - \epsilon$.

4. IMPLEMENTING SEQUENTIAL TEST

Observe that in order for the just described sequential test to recover the "non-sequential" test from Section 2.3, it suffices to utilize setup with $S = 1$ (recall that our construction fully specifies the setup component with $s = S$). This being said, with our sequential test the running time (the number of observations used to make the inference) depends on the true distribution underlying the observations, and we can use several degrees of freedom in our setup in order to save on the number of observations when the true distribution is "deeply inside" the true hypothesis. We are about to illustrate some of the options available for the basic good o.s.'s presented in Section 2.1.

4.1. Preliminaries

Assume that our o.s. is either Gaussian, or Poisson, or Discrete, see Section 2.1. We denote by n the dimension of the associated parameter vector μ: $\mu = [\mu_1; \ldots; \mu_n]$.

Assume that we are given the risk level $\epsilon \in (0, 1)$ and a collection of J nonempty compact convex sets $X_j \subset \mathcal{M}$ painted in $I \geqslant 2$ colors (i.e., the set of indices $\{1, \ldots, J\}$ is split into $I \geqslant 2$ non-overlapping nonempty sets $\mathcal{J}_1, \ldots, \mathcal{J}_I$, i being the common color of all sets X_j, $j \in \mathcal{J}_i$). Sets X_j give rise to the sets $\mathcal{X}_j$ of probability distributions defined by the corresponding densities $p_\mu(\cdot)$, $\mu \in X_j$. Note that for all considered o.s.'s different values of the parameter $\mu \in \mathcal{M}$ correspond to different probability densities p_μ. We assume that the sets X_j, $X_{j'}$ of different colors do not intersect thus giving rise to non-intersecting sets of distributions $\mathcal{X}_j$ and $\mathcal{X}_{j'}$. Given this input, we intend to specify the setup for sequential test deciding on the associated hypotheses H_i, $1 \leqslant i \leqslant I$.

Let $\psi(\mu, \nu) : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ be the rate function

$$\psi(\mu, \nu) = \ln\left(\int_\Omega \sqrt{p_\mu(\omega) p_\nu(\omega)} P(d\omega)\right),$$

associated with the o.s. in question. An immediate computation shows that (cf. [1, Section 2.3])

$$\psi(\mu, \nu) = \begin{cases} -\dfrac{1}{8}\|\mu - \nu\|_2^2,\ \ \mu, \nu \in \mathcal{M} = \mathbb{R}^n, & \text{Gaussian case} \\ -\dfrac{1}{2}\displaystyle\sum_{\ell=1}^n (\sqrt{\mu_\ell} - \sqrt{\nu_\ell})^2,\ \ \mu, \nu \in \mathcal{M} = \{x \in \mathbb{R}^n : x > 0\}, & \text{Poisson case} \\ \ln\left(\displaystyle\sum_{\omega=1}^n \sqrt{\mu_\omega \nu_\omega}\right),\ \ \mu, \nu \in \mathcal{M} = \left\{x \in \mathbb{R}^n : x > 0, \displaystyle\sum_{\omega=1}^n x_\omega = 1\right\}, & \text{Discrete case.} \end{cases}$$

Theorem states that for the considered o.s.'s, given two nonempty convex compact subsets X, Y of $\mathcal{M}$ and setting

$$\Psi_{XY} = \max_{\mu\in X,\nu\in Y} \psi(\mu,\nu),$$

the quantity $\exp\{\Psi_{XY}\}$ is exactly the risk of the detector $\phi(\cdot)$ yielded by theorem as applied to X, Y and the o.s. in question. Besides this, $\psi(\mu,\nu)$ clearly is a smooth concave and symmetric ($\psi(\mu,\nu) = \psi(\nu,\mu)$) function on its domain, and $\psi(\mu,\nu) < 0$ whenever $\mu \neq \nu$.

For $j \in \{1,\dots,J\}$, we set

$$\psi_j(\mu) = \max_{\nu\in X_j} \psi(\mu,\nu) : \mathcal{M} \to \mathbb{R}.$$

Since $\psi(\mu,\nu)$ is concave in $\mu,\nu \in \mathcal{M}$, and X_j is a convex compact set, the functions $\psi_j(\cdot)$ are concave and continuous on $\mathcal{M}$.

Let now $\mathcal{O}$ be the set of all ordered pairs (j,j'), $1 \leqslant j,j' \leqslant J$, *with j,j' of different colors;* note that $(j,j') \in \mathcal{O}$ if and only if $(j',j) \in \mathcal{O}$.

For a pair j,j', $1 \leqslant j,j' \leqslant J$, let

$$\psi_{jj'} = \Psi_{X_jX_{j'}} := \max_{\mu\in X_j,\nu\in X_{j'}} \psi(\mu,\nu) = \max_{\mu\in X_j} \psi_{j'}(\mu). \tag{11}$$

Note that for $(j,j') \in \mathcal{O}$ the convex compact sets X_j, $X_{j'}$ do not intersect, thus $\psi_{jj'} < 0$, and the objective in the optimization problem on the right hand side of (11) is negative on the compact feasible set of the problem. We set

$$d = \min_{(j,j')\in\mathcal{O}} [-\psi_{jj'}], \tag{12}$$

so that $d > 0$.

Finally, for $(j,j') \in \mathcal{O}$ and a nonnegative r we say that a linear inequality $\ell(\mu) \leqslant 0$, $\mu \in \mathcal{M}$, defines a $(jj'r)$-*cut* if for all $\mu \in X_j$ such that $\ell(\mu) \leqslant 0$,

$$\psi_{jj'}(\mu) \leqslant -r.$$

Example: default cuts. For $(j,j') \in \mathcal{O}$, let $(\mu_{jj'},\nu_{jj'})$ be (μ,ν)-components of an optimal solution to the optimization problem (11). Setting

$$e_{jj'} = \nabla_\mu \psi(\mu_{jj'},\nu_{jj'}), \quad f_{jj'} = \nabla_\nu \psi(\mu_{jj'},\nu_{jj'}), \quad (j,j') \in \mathcal{O},$$

and invoking optimality conditions for (11) along with concavity of $\psi(\cdot,\cdot)$, we get for all $(j,j') \in \mathcal{O}$:

$$\forall(\mu \in X_j, \nu \in X_{j'}) : \begin{cases} e_{jj'}^{\mathrm{T}}[\mu - \mu_{jj'}] \leqslant 0 & \text{(a)} \\ f_{jj'}^{\mathrm{T}}[\nu - \nu_{jj'}] \leqslant 0 & \text{(b)} \\ \psi(\mu,\nu) \leqslant \psi_{jj'} + e_{jj'}^{\mathrm{T}}[\mu - \mu_{jj'}] + f_{jj'}^{\mathrm{T}}[\nu - \nu_{jj'}]. & \text{(c)} \end{cases} \tag{13}$$

We conclude that setting

$$\ell_{jj'}(\mu) = \psi(\mu_{jj'},\nu_{jj'}) + e_{jj'}^{\mathrm{T}}(\mu - \mu_{jj'}) - r,$$

we get an affine function of $\mu \in \mathcal{M}$ which upper-bounds $\psi_{jj'}(\cdot) - r$ on X_j.

Indeed, for any $\nu \in X_{j'}, \mu \in X_j$

$$\begin{aligned} \psi(\mu,\nu) &\leqslant \psi(\mu_{jj'},\nu_{jj'}) + e_{jj'}^{\mathrm{T}}(\mu - \mu_{jj'}) + f_{jj'}^{\mathrm{T}}(\nu - \nu_{jj'}) \\ &\leqslant \psi(\mu_{jj'},\nu_{jj'}) + e_{jj'}^{\mathrm{T}}(\mu - \mu_{jj'}) \end{aligned}$$

(we have used (13.*c*,*b*)). Taking in the resulting inequality the supremum over $\nu \in X_{j'}$, we arrive at $\ell_{jj'}(\mu) \geqslant \psi_{j'}(\mu) - r$, $\mu \in X_j$.

The bottom line is that $\ell_{jj'}(\mu) \leqslant 0$ is a $(jj'r)$-cut; we shall refer to this cut as *default.*[6]

[6] In the long version of this paper we also describe different cuts, referred to as *smart cuts*, which aim to minimize the volume of the sets $\{\mu \in X_j,\ \psi_{jj'}(\mu) > -r\}$ (see [19, Section 4.4]).

4.2. Specifying the Setup

The setup for our sequential test is as follows.

(1) We select a sequence of positive integers $\{\bar{k}(s)\}_{s=1}^{\infty}$ satisfying

$$\bar{k}(1)=1,\ \bar{k}(s)<\bar{k}(s+1)\leqslant 2\bar{k}(s),\ s=1,2,\dots, \tag{14}$$

and specify S as the smallest positive integer such that

$$\bar{k}(S)>d^{-1}\ln\left(SJ^2/\epsilon\right) \tag{15}$$

(S is well defined due to $\bar{k}(s)\geqslant s$).

For $1\leqslant s\leqslant S$, we set

$$\epsilon_s=\frac{\epsilon}{2S},\ r(s)=\bar{k}(s)^{-1}\ln(SJ^2/\epsilon),\ \ \delta_s=\exp\{-r(s)\}. \tag{16}$$

(2) For every $j\in\{1,\dots,J\}$ and every $s\in\{1,\dots,S\}$, we specify closed convex subsets $X_{j\iota s}$, $1\leqslant\iota\leqslant\iota_{js}$, of X_j as follows. For every pair $(j,j')\in\mathcal{O}$, we select somehow a $(jj'r(s))$-cut $\ell_{jj's}(\cdot)\leqslant 0$ and set

$$\begin{aligned}X_{js}^{j'}&=\{\mu\in X_j:\ell_{jj's}(\mu)\geqslant 0,\ j'\in\overline{\mathcal{J}}_j,\}\\ X_{js}^{j}&=\{\mu\in X_j:\ell_{jj's}(\mu)\leqslant 0,\ j'\in\overline{\mathcal{J}}_j\},\end{aligned} \tag{17}$$

where for $1\leqslant j\leqslant J$ the set $\overline{\mathcal{J}}_j$ contains all indices $1\leqslant j'\leqslant J$ of the color different from that of j. Eliminating form this list all sets which are empty, we end up with a number $\iota_{js}\leqslant J$ of *nonempty* convex compact sets $X_{j\iota s}$, $1\leqslant\iota\leqslant\iota_{js}$, with X_j being their union.

Observe that $r(S)<d$ by (15), and for $(j,j')\in\mathcal{O}$ we clearly have

$$\max_{\mu\in X_j}\psi_{j'}(\mu)=\psi_{jj'}\leqslant\max_{(j,j')\in\mathcal{O}}\psi_{jj'}=-d$$

implying that $\ell_{jj'S}(\mu)\equiv -1$ are legitimate $(jj'S)$-cuts. These are exactly the cuts we use when $s=S$.

We claim that the just defined entities form a legitimate setup for a sequential test. All we need in order to justify this claim is to verify that the S-component of our setup is as required, that is, that (a) for every $j\in\{1,\dots,J\}$, $\iota_{jS}=1$, whence $X_{j1S}=X_j$ and $L_S=J$, and that (b) $q,q'\in\{1,\dots,L_S\}^2$ (recall that $L_S=J$) are S-close if and only if q and q' are of the same color.

To verify (a), note that $\ell_{jj'S}(\cdot)\equiv -1$ whenever $(j,j')\in\mathcal{O}$, implying that $X_{jS}^{j'}=\emptyset$ when $j'\in\overline{\mathcal{J}}_j$ and $X_{jS}^{j}=X_j$, as claimed in (a). To verify (b), note that as it was already mentioned, for j, j' of different colors, the risk of the detector yielded by theorem as applied to the sets $X=X_j$, $Y=X_{j'}$, is $\exp\{\psi_{jj'}\}$, that is, this risk is $\leqslant\exp\{-d\}$. Invoking the already verified (a), we conclude that $\epsilon_{qq',S}\leqslant\exp\{-d\}$ whenever $1\leqslant q,q'\leqslant L_S=J$ and q,q' are of different colors. As we have seen, $r(S)<d$, whence $\delta_S=\exp\{-r(S)\}>\exp\{-d\}$. The bottom line is that whenever $1\leqslant q,q'\leqslant J$ and q,q' are of different colors, we have $\epsilon_{qq',S}<\delta_S$; this observation combines with the definition of S-closeness to imply that q,q' are S-close if and only if q,q' are of different colors, as claimed in (b).

The legitimate setup we have presented induces a sequential test, let it be denoted by $\mathcal{T}$. We are about to analyse the properties of this test.

4.3. Analysis

Our first observation is that for the sequential setup we have presented one has $k(s) \leqslant \bar{k}(s)$, $1 \leqslant s \leqslant S$. To verify this claim, we need to check that when $k = \bar{k}(s)$, we have

$$\mathrm{Opt}(k, s) < \epsilon_s = \frac{\epsilon}{2S}.$$

As we have already mentioned (see the comment after the definition (9) of $\mathrm{Opt}(\cdot,\cdot)$), $\mathrm{Opt}(k, s)$ is the spectral norm of the entrywise nonnegative symmetric matrix D^{ks} of the size $L_s \times L_s$ with entries not exceeding δ_s^k. Since, by construction, $L_s \leqslant J^2$ the spectral norm of D^{ks} does not exceed $J^2\delta_s^k$. The latter quantity indeed is $< \epsilon_s$ when $k = \bar{k}(s)$, see (16).

Worst-case performance. In the analysis to follow, we assume that $\epsilon \in (0, \frac{1}{4})$.

By Proposition 2, the sequential test $\mathcal{T}$ always accepts at most one of the hypotheses $H_1, \ldots, H_I$, and the probability not to accept the true hypothesis is at most ϵ; moreover, the number of observations used by $\mathcal{T}$ never exceeds $k(S) \leqslant \overline{K} := \bar{k}(S)$. On the other hand, from the definition (12) of d and Corollary 1 it follows that in order for a whatever test to decide on the hypotheses $H_1, \ldots, H_I$ with risk ϵ via stationary repeated observations, the size of the observation sample should be *at least*[7]

$$K^+ = \left[\frac{\frac{1}{2}\ln(1/\epsilon) - \ln(2)}{\ln(1/\epsilon)}\right]\frac{\ln(1/\epsilon)}{d} \geqslant \frac{\ln(1/\epsilon)}{4d}. \tag{18}$$

As a result, *unless d is "astronomically small," $\overline{K}$ is within logarithmic factor of K^+*, implying quasi-optimal worst-case performance of the test $\mathcal{T}$. The precise statement is as follows:

Proposition 3. *Let $d > 0$, $J \geqslant 2$ and $\epsilon \in (0, \frac{1}{4})$ satisfy, for some $\kappa \geqslant 1$, the relation*

$$\ln(1/d) \leqslant \kappa \ln(J^2/\epsilon). \tag{19}$$

Then

$$\overline{K} \leqslant \max\left[1.5\frac{\kappa\ln(J^2/\epsilon)}{d}\right]. \tag{20}$$

For proof, see the appendix.

For all practical purposes we can assume that $d \geqslant 10^{-6}$, otherwise the *lower* bound K^+ on the number of observations required by $(1-\epsilon)$-reliable test would be impractically large. Assuming $d \geqslant 10^{-6}$, (19) is satisfied with $\kappa = 5$ (recall that $\epsilon \leqslant \frac{1}{4}$ and $J \geqslant 2$). Thus, for all practical purposes we may treat the quantity κ from the premise of Proposition 3 as a moderate absolute constant, implying that the upper bound $\overline{K}$ on the worst-case observation time of our $(1-\epsilon)$-reliable sequential test $\mathcal{T}$ indeed is within a logarithmic factor $O(1)\frac{\ln(J/\epsilon)}{\ln(1/\epsilon)}$ of the lower bound K^+ on the worst-case observation time of an "ideal" $(1-\epsilon)$-reliable test.

Remark. It is easily seen that when $\bar{k}(s)$ grows with s as rapidly as allowed by (14): $\bar{k}(s) = 2^{s-1}$, the result completely similar to the one of Proposition 3 holds true in a much wider than (19) range of values of d, specifically, in the range $\ln(1/d) \leqslant CJ^2/\epsilon$, for a whatever constant $C \geqslant 1$; and in this range, one has $\overline{K} \leqslant C'\max[1, \frac{\ln(J^2/\epsilon)}{d}]$, with C' depending solely on C.

Actual performance. For $\mu \in X := \bigcup_{j=1}^J X_j$ let $s_*(\mu)$ be the smallest $s \leqslant S$ such that for some $j \leqslant J$ it holds $\mu \in X_{js}^j$ (see (17)). Equivalently:

$$s_*(\mu) = \min\left\{s : \exists j \leqslant J : \mu \in X_j \ \&\ \ell_{jj's}(\mu) \leqslant 0 \ \forall j' \in \overline{\mathcal{J}}_j.\right\} \tag{21}$$

[7] Indeed, this is exactly the smallest number of observations which is necessary, according to Corollary 1, to separate with the risk $\leqslant \epsilon$ the pair of hypotheses corresponding to $(j, j') \in \mathcal{O}$ for which $\psi_{jj'} = -d$.

Note that $s_*(\mu)$ is well defined—we have already seen that $X_{jS}^{j'} = \emptyset$ whenever $j' \in \overline{\mathcal{J}}_j$, so that $s = S$ is feasible for the right hand side problem in (21).

Let us denote $\mathcal{X}_{js}^{\iota}$ ($\mathcal{Z}_{qs}$) the set of densities p_μ such that $\mu \in X_{js}^{\iota}$ ($\mu \in Z_{qs} = X_{js}^{\iota}$), and let $Q_s(\mu)$ be the set of the indices q of all sets $Z_{qs} \ni p_\mu$. For $\mu \in \bigcup_{j=1}^{J} X_j$, we define $s[\mu] := s[p_\mu]$, i.e., $s[\mu]$ is the smallest s such that there is a set, say, Z_{q_*s} containing μ such that all sets $\mathcal{Z}_{qs}$ of distributions p_ν such that $\nu \in Z_{qs}$ which are s-close to $\mathcal{Z}_{q_*s}$ are of the same color.

Proposition 4. *One has*

$$s[\mu] \leqslant s_*(\mu).$$

Assume that the observations are drawn from density $p_\mu(\cdot)$, $\mu \in \bigcup_j X_j$. By Proposition 2, with p_μ-probability $\geqslant 1-\epsilon$ our sequential test $\mathcal{T}$ terminates in no more than $s[\mu] \leqslant s_*(\mu)$ steps and upon termination recovers correctly the color $i[\mu]$ of μ (i.e., accepts the true hypothesis $H_{i[\mu]}$, and only this hypothesis). Thus, if μ is "deeply inside" one of the sets X_j, meaning that $s_*(\mu)$ is much smaller than S, our sequential test will, with reliability $1-\epsilon$, identify correctly the true hypothesis $H_{i[\mu]}$ much faster than in S stages.

4.4. Application in the Gaussian Case

Upper bounding $s[\mu]$. In the Gaussian case *with default cuts*, the quantity $s_*(\mu)$ which, as we have seen in Section 4.3, is a $(1-\epsilon)$-reliable upper bound on the number of stages in which $\mathcal{T}$ recognizes $(1-\epsilon)$-reliably the true hypothesis $H_{i[\mu]}$ provided the observations are drawn from $p_\mu(\cdot)$, admits a transparent geometric upper bound. Observe, first, that in the Gaussian case we have

$$d = \frac{1}{8} \min_{(j,j')\in\mathcal{O}} \min_{a\in X_j, b\in X_{j'}} \|a-b\|_2^2.$$

Now let $\rho(\mu)$ be the largest ρ such the $\|\cdot\|_2$-ball of radius ρ centered at μ is contained in certain X_j. We claim that

$$s[\mu] \leqslant s_*(\mu) \leqslant \bar{s}(\mu) := \min\left\{s \leqslant S : r(s) \leqslant d + \sqrt{d/2}\rho(\mu)\right\}, \tag{22}$$

meaning that the deeper μ is "inside" one of X_j (the larger is $\rho(\mu)$), the smaller is the number of observations needed for $\mathcal{T}$ to identify correctly the color of μ.

Justification of (22) is as follows. The first inequality in (22) is nothing but Proposition 4. To prove the second inequality, observe, first, that $\bar{s} := \bar{s}(\mu)$ is well defined (indeed, as we have seen, $r(S) < d$). Let j_* be such that the $\|\cdot\|_2$-ball B of radius $\rho(\mu)$ centered at μ is contained in X_{j_*}, and let $j \in \overline{\mathcal{J}}_{j_*}$. By (13.$a$), we have $e_{j_*j}^{\mathrm{T}}[\mu' - a_{j_*j}] \leqslant 0$ for all $\mu' \in X_{j_*}$ and thus for all $\mu' \in B$, and therefore $e_{j_*j}^{\mathrm{T}}[\mu - a_{j_*j}] \leqslant -\rho(\mu)\|e_{j_*j}\|_2$, whence for all $j \in \overline{\mathcal{J}}_{j_*}$ we have

$$\psi_{j_*j} + e_{j_*j}^{\mathrm{T}}[\mu - a_{j_*j}] \leqslant \psi_{j_*j} - \rho(\mu)\|e_{j_*j}\|_2. \tag{23}$$

Denoting by $(\bar{\mu}, \bar{\nu})$ an optimal solution to the problem $\min_{\mu\in X_{j_*}, \nu\in X_j} \|\mu-\nu\|_2$, we clearly have $\psi_{j_*j} = -\|\bar{\mu}-\bar{\nu}\|_2^2/8$, $\|e_{j_*j}\|_2 = \|\bar{\mu}-\bar{\nu}\|_2/4$, and $\|\bar{\mu}-\bar{\nu}\|^2/8 \geqslant d$ whenever $j \in \overline{\mathcal{J}}_{j_*}$ by the origin of d. Thus, (23) implies that

$$\forall (j \in \overline{\mathcal{J}}_{j_*}) : \ \psi_{j_*j} + e_{j_*j}^{\mathrm{T}}[\mu - a_{j_*j}] \leqslant -d - \rho(\mu)\sqrt{d/2},$$

that is, $\mu \in X_{j_*\bar{s}}^{j_*}$ due to $r(\bar{s}) \leqslant d + \rho(\mu)\sqrt{d/2}$, what implies $\bar{s} \geqslant s_*(\mu)$, as claimed.

Numerical illustration. The following numerical experiment highlights the power of sequential testing in the Gaussian case with default cuts. In this experiment, we are given $J = 4$ sets $X_j \subset \mathbb{R}^2$; X_1 is the square $\{0.01 \leqslant x_1, x_2 \leqslant 1\}$, X_2, X_3, X_4 are obtained from X_1 by reflections w.r.t. the coordinate axes and the origin. The partition of the index set into groups $\mathcal{J}_i$ is trivial—these groups are the elements of $\mathcal{J} = \{1, 2, 3, 4\}$, so that our goal is to recognize which of the sets X_j contains the mean μ of the observation. Our results are as follows: in the experiment with $S = 20$, $\bar{k}(s) = 2^{s-1}$, and $d = 5.0\mathrm{e} - 5$, when selecting μ in $\bigcup_{j=1}^4 X_j$ at random according to the uniform distribution, the empirical average of the number of observations before termination, which is necessary to obtain a 0.99-reliable test, is as large as 1.6×10^5. This reflects the fact that X_j are pretty close to each other. At the same time, the median number of observations before termination is just 154, reflecting the fact that in our experiment μ, with reasonably high probability, indeed is "deeply inside" the set X_j containing μ.

ACKNOWLEDGMENTS

A.B. Juditsky acknowledges the support of the CNRS-Mastodons project GARGANTUA, and the LabEx PERSYVAL-Lab (ANR-11-LABX-0025), and A.S. Nemirovski acknowledges the support of the NSF, grants CMMI-1232623, CMMI-1262063, CCF-1415498.

REFERENCES

1. Goldenshluger, A., Juditski, A., and Nemirovski, A., *Hypothesis Testing by Convex Optimization*, *arXiv preprint*, arXiv:1311.6765, 2013.
2. Barnard, G.A., Sequential Tests in Industrial Statistics, *Supplement J. Royal Statist. Soc.*, 1946, vol. 8, pp. 1–26.
3. Wald, A., Sequential Tests of Statistical Hypotheses, *Ann. Math. Statist.*, 1945, vol. 16, no. 2, pp. 117–186.
4. Wald, A., and Wolfowitz, J., Optimum Character of the Sequential Probability Ratio Test, *Ann. Math. Statist.*, 1948, vol. 19, no. 3, pp. 326–339.
5. Chernoff, H., *Sequential Analysis and Optimal Design*, SIAM, 1972, vol. 8.
6. Ghosh, B.K., A Brief History of Sequential Analysis, in *Handbook of Sequential Analysis*, New York: Marcel Dekker, 1991, pp. 1–19.
7. Bakeman, R., *Observing Interaction: An Introduction to Sequential Analysis*, Cambridge: Cambridge Univ. Press, 1997.
8. Lai, T.L., Sequential Analysis: Some Classical Problems and New Challenges, *Statist. Sinica*, 2001, vol. 11, no. 2, pp. 303–350.
9. Juditsky, A.B., and Nemirovski, A.S., Nonparametric Estimation by Convex Programming, *Ann. Statist.*, 2009, vol. 37, no. 5a, pp. 2278–2300.
10. Burnashev, M., On the Minimax Detection of an Imperfectly Known Signal in a White Noise Background, *Theory Probab. Appl.*, 1979, vol. 24, pp. 107–119.
11. Burnashev, M., Discrimination of Hypotheses for Gaussian Measures and a Geometric Characterization of the Gaussian Distribution, *Math. Notes.*, 1982, vol. 32, pp. 757–761.
12. Ingster, Y., and Suslina, I.A., *Nonparametric Goodness-of-fit Testing under Gaussian Models*, Lecture Notes Statist., vol. 169, Berlin: Springer, 2002.
13. Le Cam, L., Convergence of Estimates under Dimensionality Restrictions, *Ann. Statist.*, 1973, pp. 38–53.
14. Le Cam, L., On Local and Global Properties in the Theory of Asymptotic Normality of Experiments, *Stochast. Processes Related Topics*, 1975, vol. 1, pp. 13–54.

15. Birgé, L., Approximation dans les spaces métriques et théorie de l'estimation: Inégalités de Cràmer-Chernoff et théorie asymptotique des tests, *PhD Dissertation*, Université Paris VII, 1980.

16. Birgé, L., Sur un théorème de minimax et son application aux tests, *Probab. Math. Stat.*, 1982, vol. 3, pp. 259–282.

17. Birgé, L., Robust Testing for Independent Non Identically Distributed Variables and Markov Chains, in *Specif. Statist. Models*, Berlin: Springer, 1983, pp. 134–162.

18. Le Cam, L., *Asymptotic Methods in Statistical Decision Theory*, Series in Statistics, Berlin: Springer, 1986.

19. Juditsky, A.B., and Nemirovski, A.S., *On Sequential Hypotheses Testing via Convex Optimization*, *arXiv preprint*, arXiv:1412.1605, 2014.

This paper was recommended for publication by P.S. Shcherbakov, a member of the Editorial Board