Hypotheses testing by convex optimization

Alexander Goldenshluger * Anatoli Juditsky † Arkadi Nemirovski ‡

December 2, 2014

Abstract

We discuss a general approach to hypothesis testing. The main "building block" of the proposed construction is a test for a pair of hypotheses in the situation where each particular hypothesis states that the vector of parameters identifying the distribution of observations belongs to a convex compact set associated with the hypothesis. This test, under appropriate assumptions, is *provably nearly optimal* and is yielded by a solution to a convex optimization problem, so that the construction admits computationally efficient implementation. We further demonstrate that our assumptions are satisfied in several important and interesting applications. Finally, we show how our approach can be applied to a rather general testing problems encompassing several classical statistical settings.

1 Introduction

In this paper we promote a unified approach to a class of decision problems, based on Convex Programming. Our main building block (which we believe is important by its own right) is a construction, based on Convex Programming (and thus computationally efficient) allowing, under appropriate assumptions, to build a *provably nearly optimal* test for deciding between a pair of composite hypotheses on the distribution of observed random variable. Our approach is applicable in several important situations, primarily, those when observation (a) comes from Gaussian distribution on $\mathbb{R}^m$ parameterized by its expectation, the covariance matrix being once for ever fixed, (b) is an m-dimensional vector with independent Poisson entries, parameterized by the collection of intensities of the entries, (c) is a randomly selected point from a given m-point set $\{1, ..., m\}$, with the straightforward parametrization of the distribution by the vector of probabilities for the observation to take values 1,..., m, (d) comes from a "direct product of the outlined observation schemes," e.g., is a collection of K independent realizations of a random variable described by (a)-(c). In contrast to rather restrictive assumptions on the families of distributions we are able to handle, we are very flexible as far as the hypotheses are concerned: all we require from a hypothesis is to correspond to a convex and compact set in the "universe" $\mathcal{M}$ of parameters of the family of distributions we are working with.

*Department of Statistics, University of Haifa, 31905 Haifa, Israel, goldensh@stat.haifa.ac.il

†LJK, Université Grenoble Alpes, B.P. 53, 38041 Grenoble Cedex 9, France, anatoli.juditsky@imag.fr

‡Georgia Institute of Technology, Atlanta, Georgia 30332, USA, nemirovs@isye.gatech.edu

Research of the first author was supported by grants BSF 2010466, and ISF 104/11. The second author was supported by the CNRS-Mastodons project GARGANTUA, and the LabEx PERSYVAL-Lab (ANR-11-LABX-0025). Research of the third author was supported by NSF grants CMMI-1232623, CMMI-1262063, CCF-1415498.

As a consequence, the spirit of the results to follow is quite different from that of a "classical" statistical inquiry, where one assumes that the signals underlying noisy observations belong to some "regularity classes" and the goal is to characterize analytically the minimax rates of detection for those classes. With our approach allowing for highly diverse hypotheses, an attempt to describe analytically the quality of a statistical routine seems to be pointless. For instance, in the two-hypotheses case, all we know in advance is that the test yielded by our construction, assuming the latter applicable, is provably nearly optimal, with explicit specification of what "nearly" means presented in Theorem 2.1.ii. By itself, this "near optimality" usually is not all we need — we would like to know what actually are the performance guarantees (say, probability of wrong detection, or the number of observations sufficient to make an inference satisfying given accuracy and/or reliability specifications). The point is that with our approach, rather detailed information of this sort can be obtained by efficient situation-oriented computation. In this respect our approach follows the one of [35, 36, 7, 9, 11, 37] where what we call below "simple tests" were used to test composite hypotheses represented by convex sets of distributions[1]; later this approach was successfully applied to nonparametric estimation of signals and functionals [10, 18, 19, 12]. On the other hand, what follows can be seen as a continuation of another line of research focusing on testing [14, 15, 31] and on a closely related problem of estimating linear functionals [29, 30, 17] in white noise model. In the present paper we propose a general framework which mirrors that of [32]. Here the novelty (to the best of our understanding, essential) is in applying techniques of the latter paper to hypotheses testing rather than to estimating linear forms, which allows to naturally encompass and extend the aforementioned approaches to get provably good tests for observations schemes mentioned in (a) – (d). We strongly believe that this approach allows to handle a diverse spectrum of applications, and in this paper our focus is on efficiently implementable testing routines[2] and related elements of the "calculus of tests".

The contents and organization of the paper are as follows. We start with near-optimal testing of pairs of hypotheses, both in its general form and for particular cases of (a) – (d) (section 2). We then demonstrate (section 3) that our tests (same as other tests of similar structure) for deciding on *pairs* of hypotheses are well suited for "aggregation," via Convex Programming and simple Linear Algebra, into tests with efficiently computable performance guarantees deciding on $M \geq 2$ composite hypotheses. In the concluding section 4 our focus is on applications. Here we illustrate the implementation of the approaches developed in the preceding sections by building models and carrying out numerical experimentation for several statistical problems including Positron Emission Tomography, detection and identification of signals in a convolution model, Markov chain related inferences, and some others.

In all experiments optimization was performed using `Mosek` optimization software [1]. The proofs missing in the main body of the paper can be found in the appendix.

2 Situation and Main result

In the sequel, given a parametric family $\mathcal{P} = \{P_\mu, \mu \in \mathcal{M}\}$ of probability distributions on a space Ω and an observation $\omega \sim P_\mu$ with unknown $\mu \in \mathcal{M}$, we intend to test some composite hypotheses

[1] These results essentially cover what in the sequel is called "Discrete case," see section 2.3 for more detailed discussion.

[2] For precise definitions and details on efficient implementability, see, e.g., [6]. For the time being, it is sufficient to assume that the test statistics can be computed by a simple Linear Algebra routine with parameters which are optimal solutions to an optimization problem which can be solved using `CVX` [24].

about the parameter μ. In the situation to be considered in this paper, provably near-optimal testing reduces to Convex Programming, and we start with describing this situation.

2.1 Assumptions and goal

In what follows, we make 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
   - $p_\mu(\omega)$ 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 $p_\mu(\omega) \le p^M(\omega)$ 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*.

Now suppose that, on the top of a good observation scheme, we are given two nonempty convex compact sets $X \subset \mathcal{M}$, $Y \subset \mathcal{M}$. Given an observation $\omega \sim P_\mu$ with some unknown $\mu \in \mathcal{M}$ *known to belong either to* X (hypothesis H_X) *or to* Y (hypothesis H_Y), our goal is to decide which of the two hypotheses takes place. Let $T(\cdot)$ be a test, i.e. a Borel function on Ω taking values in $\{-1, 1\}$, which receives on input an observation ω (along with the data participating in the description of H_X and H_Y). Given observation ω, the test accepts H_X and rejects H_Y when $T(\omega) = 1$, and accepts H_Y and rejects H_X when $T(\omega) = -1$. The quality of the test is characterized by its error probabilities – the probabilities of rejecting erroneously each of the hypotheses:

$$\epsilon_X = \sup_{x \in X} P_x\{\omega : T(\omega) = -1\}, \quad \epsilon_Y = \sup_{y \in Y} P_y\{\omega : T(\omega) = 1\},$$

and we define the *risk of the test* as the maximal error probability: $\max\{\epsilon_X, \epsilon_Y\}$.

In the sequel, we focus on *simple* tests. By definition, a simple test is specified by a *detector* $\phi(\cdot) \in \mathcal{F}$; it accepts H_X, the observation being ω, if $\phi(\omega) \ge 0$, and accepts H_Y otherwise. We define the *risk* of a detector ϕ on (H_X, H_Y) as the smallest ϵ such that

$$\int_\Omega \exp\{-\phi(\omega)\}p_x(\omega)P(d\omega) \le \epsilon \ \forall x \in X, \ \int_\Omega \exp\{\phi(\omega)\}p_y(\omega)P(d\omega) \le \epsilon \ \forall y \in Y. \tag{1}$$

For a simple test with detector ϕ we have

$$\epsilon_X = \sup_{x \in X} P_x\{\omega : \phi(\omega) < 0\}, \quad \epsilon_Y = \sup_{y \in Y} P_y\{\omega : \phi(\omega) \ge 0\},$$

and the risk $\max\{\epsilon_X, \epsilon_Y\}$ of such test clearly does not exceed the risk ϵ of the detector ϕ.

2.2 Main result

We are about to show that in the situation in question, *an efficiently computable via Convex Programming detector results in a nearly optimal test.* The precise statement is as follows:

Theorem 2.1 *In the just described situation and under the above assumptions,*

(i) *The function*

$$\Phi(\phi,[x;y]) = \ln\left(\int_\Omega \exp\{-\phi(\omega)\}p_x(\omega)P(d\omega)\right) + \ln\left(\int_\Omega \exp\{\phi(\omega)\}p_y(\omega)P(d\omega)\right) : \mathcal{F}\times(X\times Y)\to\mathbb{R}. \tag{2}$$

*is continuous on its domain, is convex in $\phi(\cdot)\in\mathcal{F}$, concave in $[x;y]\in X\times Y$, and possesses a saddle point (*min *in ϕ,* max *in $[x;y]$) $(\phi_*(\cdot),[x_*;y_*])$ on $\mathcal{F}\times(X\times Y)$. ϕ_* w.l.o.g. can be assumed to satisfy the relation*[3]

$$\int_\Omega \exp\{-\phi_*(\omega)\}p_{x_*}(\omega)P(d\omega) = \int_\Omega \exp\{\phi_*(\omega)\}p_{y_*}(\omega)P(d\omega). \tag{3}$$

Denoting the common value of the two quantities in (3) by $\varepsilon_\star$, the saddle point value

$$\min_{\phi\in\mathcal{F}}\max_{[x;y]\in X\times Y}\Phi(\phi,[x;y])$$

is $2\ln(\varepsilon_\star)$, and the risk of the simple test associated with the detector ϕ_* on the composite hypotheses H_X, H_Y is $\le\varepsilon_\star$. Moreover, for every $a\in\mathbb{R}$, for the test with the detector $\phi_*^a(\cdot)\equiv\phi_*(\cdot)-a$, the probabilities ϵ_X to reject H_X when the hypothesis is true and ϵ_Y to reject H_Y when the hypothesis is true can be upper-bounded as

$$\epsilon_X\le\exp\{a\}\varepsilon_\star,\ \epsilon_Y\le\exp\{-a\}\varepsilon_\star. \tag{4}$$

(ii) *Let $\epsilon\ge 0$ be such that there exists a (whatever) test for deciding between two simple hypotheses*

$$(A):\omega\sim p(\cdot):=p_{x_*}(\cdot),\quad (B):\omega\sim q(\cdot):=p_{y_*}(\cdot) \tag{5}$$

with the sum of error probabilities $\le 2\epsilon$. Then

$$\varepsilon_\star\le 2\sqrt{\epsilon(1-\epsilon)}.$$

In other words, if the simple hypotheses (A), (B) can be decided, by a whatever test, with the sum of error probabilities 2ϵ, then the risk of the simple test with detector ϕ_* on the composite hypotheses H_X, H_Y does not exceed $2\sqrt{\epsilon(1-\epsilon)}$.

(iii) *The detector ϕ_* specified in (i) is readily given by the $[x;y]$-component $[x_*;y_*]$ of the associated saddle point of Φ, specifically,*

$$\phi_*(\cdot)=\tfrac{1}{2}\ln\left(p_{x_*}(\cdot)/p_{y_*}(\cdot)\right). \tag{6}$$

[3] Note that $\mathcal{F}$ contains constants, and shifting by a constant the ϕ-component of a saddle point of Φ and keeping its $[x;y]$-component intact, we clearly get another saddle point of Φ.

Remark. At this point let us make a small summary of the properties of simple tests in the problem setting and under assumptions of section 2.1:

(i) One has

$$\varepsilon_* = \exp(\mathrm{Opt}/2) = \rho(x_*, y_*),$$

where $[x_*; y_*]$ is the $[x; y]$-component of the saddle point solution of (2), and

$$\rho(x, y) = \int_\Omega \sqrt{p_x(\omega)p_y(\omega)}P(d\omega),$$

is the *Hellinger affinity* of distributions p_x and p_y [34, 37];

(ii) the optimal detector ϕ_* as in (6) satisfies (1) with $\epsilon = \varepsilon_*$;

(iii) the simple test with detector ϕ_* can be "skewed", by using instead of $\phi_*(\cdot)$ detector $\phi_*^a(\cdot) = \phi_*(\cdot) - a$, to attain error probabilities of the test $\epsilon_X = e^a\varepsilon_*$ and $\epsilon_Y = e^{-a}\varepsilon_*$.

As we will see in an instant, the properties (i) – (iii) of simple tests allow to "propagate" the near-optimality property of the tests in the case of repeated observations and multiple testing, and underline all further developments.

Of course, the proposed setting and construction of simple test are by no means unique. For instance, any test $\overline{T}$ in the problem of deciding between H_X and H_Y, with the risk bounded with $\bar\epsilon \in (0, 1/2)$, gives rise to the detector

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

(recall that $\overline{T}(\omega) = 1$ when $\overline{T}$, as applied to observation ω, accepts H_X, and $\overline{T}(\omega) = -1$ otherwise). One can easily see that the risk of $\bar\phi(\cdot)$ satisfies the bounds of (1) with

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

In other words, in the problem of deciding upon H_X and H_Y, any test $\overline{T}$ with the risk $\leq \bar\epsilon$ brings about a simple test with detector $\bar\phi$, albeit with a larger risk ϵ.

2.3 Basic examples

We list here some situations where our assumptions are satisfied and thus Theorem 2.1 is applicable.

2.3.1 Gaussian observation scheme

In the *Gaussian observation scheme* we are given an observation $\omega \in \mathbb{R}^m$, $\omega \sim \mathcal{N}(\mu, \Sigma)$ with unknown parameter $\mu \in \mathbb{R}^m$ and known covariance matrix Σ. Here the family $\mathcal{P}$ is defined with (Ω, P) being $\mathbb{R}^m$ with the Lebesque measure, $p_\mu = \mathcal{N}(\mu, \Sigma)$, $\mathcal{M} = \mathbb{R}^m$, and $\mathcal{F} = \{\phi(\omega) = a^T\omega + b :\ a \in \mathbb{R}^m,\ b \in \mathbb{R}\}$ is the space of all affine functions on $\mathbb{R}^m$. Taking into account that

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

we conclude that Gaussian observation scheme is good. The test yielded by Theorem 2.1 is particularly simple in this case: assuming that the nonempty convex compact sets $X \subset \mathbb{R}^m$, $Y \subset \mathbb{R}^m$ do not intersect[4], and that the covariance matrix Σ of the distribution of observation is nondegenerate, we get

$$\begin{array}{l}\phi_*(\omega) = \xi^T\omega - \alpha,\ \xi = \frac{1}{2}\Sigma^{-1}[x_* - y_*],\ \alpha = \frac{1}{2}\xi^T\Sigma^{-1}[x_* + y_*],\\ \varepsilon_\star = \exp\left(-\frac{1}{8}(x_* - y_*)^T\Sigma^{-1}(x_* - y_*)\right)\\ \left[[x_*; y_*] \in \mathrm{Argmax}_{x\in X, y\in Y}\left[\psi(x,y) = -\frac{1}{4}(x-y)^T\Sigma^{-1}(x-y)\right]\right].\end{array} \tag{7}$$

One can easily verify that the error probabilities $\epsilon_X(\phi^*)$ and $\epsilon_Y(\phi^*)$ of the associated simple test do not exceed $\epsilon_* = \mathrm{Erf}\left(\frac{1}{2}\|\Sigma^{-1/2}(x_* - y_*)\|_2\right)$, where $\mathrm{Erf}(s)$ is the error function:

$$\mathrm{Erf}(t) = (2\pi)^{-1/2}\int_t^\infty \exp\{-s^2/2\}ds.$$

Moreover, in the case in question the sum of the error probabilities of our test is exactly the minimal, over all possible tests, sum of error probabilities when deciding between the simple hypotheses stating that $x = x_*$ and $y = y_*$.

Remarks. Consider the simple situation where the covariance matrix Σ is proportional to the identity matrix: $\Sigma = \sigma^2 I$ (the case of general Σ reduces to this "standard case" by simple change of variables). In this case, in order to construct the optimal test, one should find the closest in the Euclidean distance points $x_* \in X$ and $y_* \in Y$, so that the affine form $\zeta(u) = [x_* - y_*]^T u$ strongly separates X and Y. On the other hand, testing in the white Gaussian noise between the closed half-spaces $\{u : \zeta(u) \le \zeta(y_*)\}$ and $\{u : \zeta(u) \ge \zeta(x_*)\}$ (which contain Y and X, respectively) is exactly the same as deciding on two simple hypotheses stating that $y = y_*$, and $x = x_*$. Though this result is almost self-evident, it seems first been noticed in [14] in the problem of testing in white noise model, and then exploited in [15, 31] in the important to us context of hypothesis testing.

As far as numerical implementation of the testing routines is concerned, numerical stability of the proposed test is an important issue. For instance, it may be useful to know the testing performance when the optimization problem (7) is not solved to exact optimality, or when errors may be present in description of the sets X and Y. Note that one can easily bound the error of the obtained test in terms of the magnitude of violation of first-order optimality conditions for (7), which read:

$$(y_* - x_*)^T\Sigma^{-1}(x - x_*) + (x_* - y_*)^T\Sigma^{-1}(y - y_*) \le 0,\ \forall x \in X,\ y \in Y.$$

Now assume that instead of the optimal test $\phi_*(\cdot)$ we have at our disposal an "approximated" simple test associated with

$$\tilde\phi(\omega) = \tilde\xi^T\omega - \tilde\alpha,\ \tilde\xi = \frac{1}{2}\Sigma^{-1}[\tilde x - \tilde y],\ \tilde\alpha = \frac{1}{2}\tilde\xi^T[\tilde x + \tilde y],$$

where $\tilde x \in X,\ \tilde y \in Y,\ \tilde x \ne \tilde y$ satisfy

$$(\tilde y - \tilde x)^T\Sigma^{-1}(x - \tilde x) + (\tilde x - \tilde y)^T\Sigma^{-1}(y - \tilde y) \le \delta,\ \forall x \in X,\ y \in Y, \tag{8}$$

[4] otherwise $\phi_* \equiv 0$ and $\varepsilon_\star = 1$, in full accordance with the fact that in the case in question no nontrivial (i.e., with both error probabilities $< 1/2$) testing is possible.

with some $\delta > 0$. This implies the bound for the risk of the test with detector $\tilde{\phi}(\cdot)$:

$$\max[\epsilon_X, \epsilon_Y] \leq \tilde{\epsilon} = \mathrm{Erf}\left(\tfrac{1}{2}\|\Sigma^{-1/2}(\tilde{x}-\tilde{y})\|_2 - \frac{\delta}{\|\Sigma^{-1/2}(\tilde{x}-\tilde{y})\|_2}\right). \tag{9}$$

Indeed, (8) implies that $\tilde{\xi}^T(x-\tilde{x}) \geq -\frac{\delta}{2}$, $\tilde{\xi}^T(y-\tilde{y}) \leq \frac{\delta}{2}$, $\forall x \in X$, $y \in Y$. As a result,

$$\tilde{\xi}^T x - \tilde{\alpha} = \tilde{\xi}^T(x-\tilde{x}) + \tilde{\xi}^T\Sigma\tilde{\xi} \geq -\frac{\delta}{2} + \tilde{\xi}^T\Sigma\tilde{\xi} \ \ \forall x \in X.$$

and for all $x \in X$,

$$\mathrm{Prob}_x\{\tilde{\phi}(\omega) < 0\} = \mathrm{Prob}_x\{\tilde{\xi}^T(\omega - x) < -\tilde{\xi}^T x + \tilde{\alpha}\} = \mathrm{Prob}_x\left\{\|\Sigma^{1/2}\tilde{\xi}\|_2\eta < -\|\Sigma^{1/2}\tilde{\xi}\|_2^2 + \frac{\delta}{2}\right\},$$

where $\eta \sim \mathcal{N}(0,1)$. We conclude that

$$\epsilon_X = \sup_{x\in X}\mathrm{Prob}_x\{\tilde{\phi}(\omega) < 0\} \leq \mathrm{Erf}\left(\tfrac{1}{2}\|\Sigma^{1/2}\tilde{\xi}\|_2 - \frac{\delta}{2\|\Sigma^{1/2}\tilde{\xi}\|_2}\right)$$

what implies the bound (9) for ϵ_X. The corresponding bound for $\epsilon_Y = \sup_{y\in Y}\mathrm{Prob}_y\{\tilde{\phi}(\omega) \geq 0\}$ is obtained in the same way.

2.3.2 Discrete observation scheme

Assume that we observe a realization of a random variable ω taking values in $\{1, 2, ..., m\}$ with probabilities μ_i, $i = 1, ..., m$:

$$\mu_i = \mathrm{Prob}\{\omega = i\}, \ i = 1, ..., m.$$

The just described *Discrete observation scheme* corresponds to (Ω, P) being $\{1, ..., m\}$ with counting measure, $p_\mu(\omega) = \mu_\omega$, $\mu \in \mathcal{M} = \{\mu \in \mathbb{R}^m : \ \mu_i > 0, \ \sum_{i=1}^m \mu_i = 1\}$, In this case $\mathcal{F} = \mathbb{R}(\Omega) = \mathbb{R}^m$, and 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 $\mu \in \mathcal{M}$. We conclude that Discrete observation scheme is good. Furthermore, when assuming the convex compact sets $X \subset \mathcal{M}$, $Y \subset \mathcal{M}$ (recall that in this case $\mathcal{M}$ is the relative interior of the standard simplex in $\mathbb{R}^m$) not intersecting, we get

$$\begin{array}{c}\phi_*(\omega) = \ln\left(\sqrt{[x_*]_\omega/[y_*]_\omega}\right), \ \varepsilon_\star = \exp\{\mathrm{Opt}/2\} = \rho(x_*, y_*), \\ \left[[x_*; y_*] \in \mathrm{Argmax}_{x\in X, y\in Y}\left[\psi(x,y) = 2\ln\rho(x,y), \ \mathrm{Opt} = \psi(x_*, y_*)\right],\right]\end{array} \tag{10}$$

where $\rho(x,y) = \sum_{\ell=1}^m \sqrt{x_\ell y_\ell}$ is the Hellinger affinity of distributions x and y. One has $\varepsilon_\star = \rho(x_*, y_*) = 1 - h^2(x_*, y_*)$, the Hellinger affinity of the sets X and Y, where

$$h^2(x,y) = \tfrac{1}{2}\sum_{\ell=1}^m (\sqrt{x_\ell} - \sqrt{y_\ell})^2$$

is the *Hellinger distance* between distributions x and y. Thus the result of Theorem 2.1, as applied to Discrete observation model, allows for the following simple interpretation: to construct the simple test ϕ_* one should find the closest in Hellinger distance points $x_* \in X$ and $y_* \in Y$; then the risk of the likelihood ratio test ϕ_* for distinguishing x_* from y_*, as applied to our testing problem, is bounded with $\rho(x_*, y_*) = 1 - h^2(x_*, y_*)$, the Hellinger affinity of sets X and Y.

Remarks. Discrete observation scheme considered in this section is a simple particular case – that of finite Ω – of the result of [8, 9] on distinguishing convex sets of distributions. Roughly, the situation considered in those papers is as follows: let Ω be a Polish space, P be a σ-finite σ-additive Borel measure on Ω, and $p(\cdot)$ be a density w.r.t. P of probability distribution of observation ω. Note that the corresponding observation scheme (with $\mathcal{M}$ being the set of densities with respect to P on Ω) does not satisfy the premise of section 2.1 because the linear space $\mathcal{F}$ spanned by constants and functions of the form $\ln(p(\cdot)/q(\cdot))$, $p, q \in \mathcal{M}$ is not finite-dimensional. Now assume that we are given two non-overlapping convex closed subsets X, Y of the set of probability densities with respect to P on Ω. Observe that for every positive Borel function $\psi(\cdot) : \Omega \to \mathbb{R}$, the detector ϕ given by $\phi(\omega) = \ln(\psi(\omega))$ for evident reasons satisfies the relation

$$\max_{p\in X, q\in Y} \left[\int_\Omega \mathrm{e}^{-\phi(\omega)} p(\omega)P(d\omega), \int_\Omega \mathrm{e}^{\phi(\omega)} q(\omega)P(d\omega)\right] \le \epsilon,$$
$$\epsilon = \max\left[\sup_{p\in X} \int \psi^{-1}(\omega)p(\omega)P(d\omega), \sup_{q\in Y} \int \psi(\omega)q(\omega)P(d\omega)\right]$$

Let now

$$\mathrm{Opt} = \max_{p\in X, q\in Y} \left\{\rho(p,q) = \int_\Omega \sqrt{p(\omega)q(\omega)}P(d\omega)\right\}, \tag{11}$$

which is an infinite-dimensional convex program with respect to $p \in X$ and $q \in Y$. Assuming the program solvable with an optimal solution composed of distribution $p_*(\cdot)$, $q_*(\cdot)$ which are positive, and setting $\psi_*(\omega) = \sqrt{p_*(\omega)/q_*(\omega)}$, under some "regularity assumptions" (see, e.g., Proposition 4.2 of [9]) the optimality conditions for (11) read:

$$\min_{p\in X, q\in Y} \left[\int_\Omega \psi_*^{-1}(\omega)[p_*(\omega) - p(\omega)]P(d\omega) + \int_\Omega \psi_*(\omega)[q_*(\omega) - q(\omega)]P(d\omega)\right] = 0.$$

In other words,

$$\max_{p\in X} \int_\Omega \psi_*^{-1}(\omega)p(\omega)dP(\omega) \le \int_\Omega \psi_*^{-1}(\omega)p_*(\omega)dP(\omega) = \mathrm{Opt},$$

and similarly,

$$\max_{q\in Y} \int_\Omega \psi_*(\omega)q(\omega)dP(\omega) \le \int_\Omega \psi_*(\omega)q_*(\omega)dP(\omega) = \mathrm{Opt},$$

so that for our ψ_*, we have $\epsilon = \mathrm{Opt}$.

Note that, although this approach is not restricted to the Discrete case *per se*, when Ω is not finite, the optimization problem in (11) is generally computationally intractable (the optimal detectors can be constructed explicitly for some special sets of distribution, see [9, 11]).

The bound $\varepsilon_\star$ for the risk of the simple test can be compared to the *testing affinity* $\pi(X, Y)$ between X and Y,

$$\pi(X,Y) = \max_{x\in X, y\in Y} \left\{\pi(x,y) = \sum_{\ell=1}^m \min[x_\ell, y_\ell]\right\},$$

which is the least possible sum of error probabilities $\epsilon_X + \epsilon_Y$ when distinguishing between H_X and H_Y (cf. [35, 37]). The corresponding *minimax test* is a simple test with detector $\overline{\phi}(\cdot,\cdot)$, defined according to

$$\overline{\phi}(\omega) = \ln\left(\sqrt{[\overline{x}]_\omega/[\overline{y}]_\omega}\right),$$
$$\left[[\overline{x};\overline{y}] \in \mathrm{Argmax}_{x\in X, y\in Y}\left[\textstyle\sum_{\ell=1}^m \min[x_\ell, y_\ell]\,.\right]\right.$$

Unfortunately, this test cannot be easily extended to the case where repeated observations (e.g., independent realizations ω_k, $k = 1, ..., K$, of ω) are available. In [27] such an extension has been proposed in the case where X and Y are dominated by bi-alternating capacities (see, e.g., [28, 5, 13, 3], and references therein); explicit constructions of the test were proposed for some special sets of distributions [26, 42, 41]. On the other hand, as we shall see in section 2.4, the simple test $\phi_*(\cdot, \cdot)$ allows for a straightforward generalization to the repeated observations case with the same (near-)optimality guaranties as those of Theorem 2.1.ii.

Finally, same as in the Gaussian observation scheme, the risk of a simple test with detector $\tilde{\phi}(\omega) = \frac{1}{2} \ln(\tilde{x}_\omega / \tilde{y}_\omega)$, $\omega \in \Omega$, defined by a pair of distributions $[\tilde{x}; \tilde{y}] \in X \times Y$, can be assessed through the magnitude of violation by $\tilde{x}$ and $\tilde{y}$ of the first-order optimality conditions for the optimization problem in (10). Indeed, assume that

$$\sum_{\ell=1}^{m} \sqrt{\frac{\tilde{y}_\ell}{\tilde{x}_\ell}}(x_\ell - \tilde{x}_\ell) + \sum_{\ell=1}^{m} \sqrt{\frac{\tilde{x}_\ell}{\tilde{y}_\ell}}(y_\ell - \tilde{y}_\ell) \le \delta \ \ \forall x \in X, \ y \in Y.$$

We conclude that

$$\begin{aligned} \epsilon_X &\le \max_{x \in X} \sum_{\ell=1}^{m} e^{-\tilde{\phi}_\ell} x_\ell = \max_{x \in X} \sum_{\ell=1}^{m} \sqrt{\frac{\tilde{y}_\ell}{\tilde{x}_\ell}} x_\ell \le \sum_{\ell=1}^{m} \sqrt{\tilde{y}_\ell \tilde{x}_\ell} + \delta, \\ \epsilon_Y &\le \max_{y \in Y} \sum_{\ell=1}^{m} e^{\tilde{\phi}_\ell} y_\ell = \max_{y \in Y} \sum_{\ell=1}^{m} \sqrt{\frac{\tilde{x}_\ell}{\tilde{y}_\ell}} y_\ell \le \sum_{\ell=1}^{m} \sqrt{\tilde{x}_\ell \tilde{y}_\ell} + \delta, \end{aligned}$$

so that the risk of the test $\tilde{\phi}$ is bounded with $\rho(\tilde{x}, \tilde{y}) + \delta$.

2.3.3 Poisson observation scheme

Suppose that we are given m realizations of independent Poisson random variables

$$\omega_i \sim \text{Poisson}(\mu_i)$$

with parameters μ_i, $i = 1, ..., m$. The *Poisson observation scheme* is given by (Ω, P) being $\mathbb{Z}_+^m$ with counting measure, $p_\mu(\omega) = \frac{\mu^\omega}{\omega!} e^{-\sum_i \mu_i}$ where $\mu \in \mathcal{M} = \text{int}\, \mathbb{R}_+^m$, and, similarly to the Gaussian case, $\mathcal{F}$ is comprised of the restrictions onto $\mathbb{Z}_+^m$ of affine functions: $\mathcal{F} = \{\phi(\omega) = a^T \omega + b : \ a \in \mathbb{R}^m, \ b \in \mathbb{R}\}$. Since

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

is concave in μ, we conclude that Poisson observation scheme is good.

Assume now that, same as above, in the Poisson observation scheme, the convex compact sets $X \subset \mathbb{R}_{++}^m$, $Y \subset \mathbb{R}_{++}^m$ do not intersect. Then the data associated with the simple test yielded by Theorem 2.1 is as follows:

$$\begin{array}{c} \phi_*(\omega) = \xi^T \omega - \alpha, \ \xi_\ell = \frac{1}{2} \ln([x_*]_\ell / [y_*]_\ell), \ \ \alpha = \frac{1}{2} \sum_{\ell=1}^{m} [x_* - y_*]_\ell, \ \varepsilon_\star = \exp\{\text{Opt}/2\} \\ \left[\ [x_*; y_*] \ \in \ \text{Argmax}_{x \in X, y \in Y} \left[\psi(x, y) = -2h^2(x, y) \right], \ \text{Opt} = \psi(x_*, y_*), \ \right] \end{array} \tag{12}$$

where $h^2(x, y) = \frac{1}{2} \sum_{\ell=1}^{m} \left[\sqrt{x_\ell} - \sqrt{y_\ell} \right]^2$ is the Hellinger distance between $x \in \mathbb{R}_+^m$ and $y \in \mathbb{R}_+^m$.

Remark. Let $\tilde{\phi}(\omega) = \tilde{\xi}^T\omega - \tilde{\alpha}$ be a detector, generated by $[\tilde{x};\tilde{y}] \in X \times Y$, namely, such that

$$\tilde{\xi}_\ell = \tfrac{1}{2}\ln(\tilde{x}_\ell/\tilde{y}_\ell), \quad \tilde{\alpha} = \tfrac{1}{2}\sum_{\ell=1}^{m}(\tilde{x}_\ell - \tilde{y}_\ell).$$

We assume that $[\tilde{x};\tilde{y}]$ is an approximate solution to (12) in the sense that the first-order optimality condition of (12) is 'δ-satisfied":

$$\sum_{\ell=1}^{m}\left[\left(\sqrt{\tilde{y}_\ell/\tilde{x}_\ell} - 1\right)(x_\ell - \tilde{x}_\ell) + \left(\sqrt{\tilde{x}_\ell/\tilde{y}_\ell} - 1\right)(y_\ell - \tilde{y}_\ell)\right] \le \delta \ \ \forall x \in X,\ y \in Y.$$

One can easily verify that the risk of the test, associated with $\tilde{\phi}$, is bounded with $\exp(-h^2(\tilde{x},\tilde{y}) + \delta)$ (cf. the corresponding bounds for the Gaussian and Discrete observation schemes).

2.4 Repeated observations

Good observation schemes admit naturally defined direct products. To simplify presentation, we start with explaining the corresponding construction in the case of *stationary repeated observations* described as follows.

2.4.1 K-repeated stationary observation scheme

We are given a good observation scheme $((\Omega, P), \{p_\mu(\cdot) : \mu \in \mathcal{M}\}, \mathcal{F})$ and a positive integer K, along with same as above X, Y. Instead of a single realization $\omega \sim p_\mu(\cdot)$, we now observe a sample of K *independent* realizations $\omega_k \sim p_\mu(\cdot)$, $k = 1, ..., K$. Formally, this corresponds to the observation scheme with the observation space $\Omega^K = \{\omega^K = (\omega_1, ..., \omega_K) : \omega_k \in \Omega\, \forall k\}$ equipped with the measure $P^K = P \times ... \times P$, the family $\{p_\mu^K(\omega^K) = \prod_{k=1}^K p_\mu(\omega_k), \mu \in \mathcal{M}\}$ of densities of repeated observations w.r.t. P^K, and $\mathcal{F}^K = \{\phi^K(\omega^K) = \sum_{k=1}^K \phi(\omega_k), \phi \in \mathcal{F}\}$. The components X, Y of our setup are the same as for the original single-observation scheme, and the composite hypotheses we intend to decide upon state now that the K-element observation ω^K comes from a distribution $p_\mu^K(\cdot)$ with $\mu \in X$ (hypothesis H_X) or with $\mu \in Y$ (hypothesis H_Y).

It is immediately seen that the just described K-*repeated* observation scheme is good (i.e., satisfies all our assumptions), provided that the "single observation" scheme we start with is so. Moreover, the detectors ϕ_*, ϕ_*^K and risk bounds $\varepsilon_\star$, $\varepsilon_\star^{(K)}$ given by Theorem 2.1 as applied to the original and the K-repeated observation schemes are linked by the relations

$$\phi_*^K(\omega_1, ..., \omega_K) = \textstyle\sum_{k=1}^K \phi_*(\omega_k), \ \ \varepsilon_\star^{(K)} = (\varepsilon_\star)^K. \tag{13}$$

As a result, the "near-optimality claim" Theorem 2.1.ii can be reformulated as follows:

Proposition 2.1 *Assume that for some integer $\bar{K} \ge 1$ and some $\epsilon \in (0, 1/4)$, the hypotheses H_X, H_Y can be decided, by a whatever procedure utilising $\bar{K}$ observations, with error probabilities $\le \epsilon$. Then with*

$$K^+ = \left\lfloor \frac{2\bar{K}}{1 - \frac{2\ln[2]}{\ln[1/\epsilon]}} \right\rfloor$$

observations, $\lfloor a \rfloor$ being the smallest integer $\ge a$, the simple test with the detector $\phi_*^{K^+}$ decides between H_X and H_Y with risk $\le \epsilon$.

Indeed, applying (13) with $K = \bar{K}$ and utilizing Theorem 2.1.ii, we get $\varepsilon_\star \le (2\sqrt{\epsilon})^{1/\bar{K}}$ and therefore, by the same (13), $\varepsilon_\star^{(K)} = \varepsilon_\star^K \le (2\sqrt{\epsilon})^{K/\bar{K}}$ for all K. Thus, $\varepsilon_\star(K^+) \le \epsilon$, and therefore the conclusion of Proposition follows from Theorem 2.1.i as applied to observations ω^{K^+}.

We see that for small ϵ, the "suboptimality ratio" (i.e., the ratio $K^+/\bar{K}$) of the proposed test when ϵ-reliable testing is sought is close to 2 for small ϵ.

2.4.2 Non-stationary repeated observations

We are about to define the notion of a general-type direct product of good observation schemes. The situation now is as follows: we are given K good observation schemes

$$\mathcal{O}_k = ((\Omega_k, P_k), \mathcal{M}_k \subset \mathbb{R}^{m_k}, \{p_{k,\mu_k}(\cdot) : \mu_k \in \mathcal{M}_k\}, \mathcal{F}_k)\,,\ k = 1, ..., K$$

and observe a sample $\omega^K = (\omega_1, ..., \omega_K)$ of realizations $\omega_k \in \Omega_k$ drawn independently of each other from the distributions with densities, w.r.t. P_k, being $p_{k,\mu_k}(\cdot)$, for a collection $\mu^K = (\mu_1, ..., \mu_K)$ with $\mu_k \in \mathcal{M}_k$, $1 \le k \le K$. Setting

$$\begin{aligned}
&\Omega^K = \Omega_1 \times ... \times \Omega_K = \{\omega^K = (\omega_1, ..., \omega_K) : \omega_k \in \Omega_k\, \forall k \le K\},\\
&P^K = P_1 \times ... \times P_K\\
&\mathcal{M}^K = \mathcal{M}_1 \times ... \times \mathcal{M}_K = \{\mu^K = (\mu_1, ..., \mu_K) : \mu_k \in \mathcal{M}_k\, \forall k \le K\},\\
&p_{\mu^K}(\omega^K) = p_{1,\mu_1}(\omega_1)p_{2,\mu_2}(\omega_2)...p_{K,\mu_K}(\omega_K) \quad [\mu^K \in \mathcal{M}^K, \omega^K \in \Omega^K],\\
&\mathcal{F}^K = \{\phi^K(\omega^K) = \phi_1(\omega_1) + \phi_2(\omega_2) + ... + \phi_K(\omega_K) :\ \Omega^K \to \mathbb{R} : \phi_k(\cdot) \in \mathcal{F}_k\, \forall k \le K\},
\end{aligned}$$

we get an observation scheme $((\Omega^K, P_K), \mathcal{M}^k, \{p_{\mu^K}(\cdot) : \mu^k \in \mathcal{M}^k\}, \mathcal{F}^K)$ which we call the *direct product of* $\mathcal{O}_1, ..., \mathcal{O}_K$ and denote $\mathcal{O}^K = \mathcal{O}_1 \times ... \times \mathcal{O}_K$. It is immediately seen that this scheme is good. Note that the already defined stationary repeated observation scheme deals with a special case of the direct product construction, the one where all factors in the product are identical to each other, and where, in addition, we replace $\mathcal{M}^K$ with its "diagonal part" $\{\mu^K = (\mu, \mu, ..., \mu),\ \mu \in \mathcal{M}\}$.

Let $\mathcal{O}^K = \mathcal{O}_1 \times ... \times \mathcal{O}_K$, where, for every $k \le K$,

$$\mathcal{O}_k = ((\Omega_k, P_k), \mathcal{M}_k, \{p_{\mu_k}(\cdot) : \mu_k \in \mathcal{M}_k\}, \mathcal{F}_k)$$

is a good observation scheme, specifically, either Gaussian, or Discrete, or Poisson (see section 2.3). To simplify notation, we assume that all Poisson factors $\mathcal{O}_k$ are "scalar," that is, ω_k is drawn from Poisson distribution with parameter μ_k.[5] For

$$\phi^K(\omega^K) = \sum_{k=1}^K \phi_k(\omega_k) \in \mathcal{F}^K,\ \ \mu^K = (\mu_1, ..., \mu_K) \in \mathcal{M}^K,$$

[5] This assumption in fact does not restrict generality, since an m-dimensional Poisson observation scheme from section 2.3.3 is nothing but the direct product of m scalar Poisson observation schemes. Since the direct product of observation schemes clearly is associative, we always can reduce the situation with multidimensional Poisson factors to the case where all these factors are scalar ones.

let us set

$$\Psi(\phi^K(\cdot),\mu^K)=\ln\left(\int_{\Omega^K}\exp\{-\phi^K(\omega^K)\}p_{\mu^K}(\omega^K)P^K(d\omega^K)\right)=\sum_{k=1}^K\Psi_k(\phi_k(\cdot),\mu_k),$$

with

$$\Psi_k(\phi_k(\cdot),\mu_k)=\ln(\left(\int_{\Omega_k}\exp\{-\phi_k(\omega_k)\}p_{k,\mu_k}(\omega_k)P_k(d\omega_k)\right).$$

The function $\Phi(\phi^K,[x,y])$, defined by (2) as applied to the observation scheme $\mathcal{O}^K$, clearly is

$$\Phi(\phi^K,[x;y])=\sum_{k=1}^K\left[\Psi_k(\phi_k,x_k)+\Psi_k(-\phi_k,y_k)\right],$$
$$\left[\phi^K(\omega^K)=\sum_k\phi_k(\omega_k),\ x=[x_1;...;x_k]\in\mathcal{M}^K,\quad y=[y_1;...;y_K]\in\mathcal{M}^K\right]$$

so that

$$\min_{\phi^K\in\mathcal{F}^K}\Phi(\phi^K,[x;y])=\sum_{k=1}^K\psi_k(x_k,y_k),$$

where functions $\psi_k(\cdot,\cdot)$ are defined as follows (cf. (7), (10) and (12)):

- $\psi_k(\mu_k,\nu_k)=-\frac{1}{4}(\mu_k-\nu_k)^T\Sigma_k^{-1}(\mu_k-\nu_k)$ in the case of Gaussian $\mathcal{O}_k$ with $\omega_k\in\mathbb{R}^{m_k}$, $\omega_k\sim\mathcal{N}(\mu_k,\Sigma_k)$, $\mu_k,\nu_k\in\mathbb{R}^{m_k}$;
- $\psi_k(\mu_k,\nu_k)=-(\sqrt{\mu_k}-\sqrt{\nu_k})^2$ for scalar Poisson $\mathcal{O}_k$, with $\mu_k,\nu_k>0$;
- $\psi_k(\mu_k,\nu_k)=2\ln\left(\sum_{i=1}^{m_k}\sqrt{[\mu_k]_i[\nu_k]_i}\right)$ for Discrete $\mathcal{O}_k$ with $\Omega_k=\{1,...,m_k\}$, $\mu_k,\nu_k\in\mathcal{M}_k=\{\mu\in\mathbb{R}^{m_k}:\mu>0,\ \sum_i[\mu]_i=1\}$.

Let X_k and Y_k be compact convex subsets of $\mathcal{M}_k$, $k=1,...,K$; let $X=X_1\times...\times X_K$ and $Y=Y_1\times...\times Y_K$. Assume that $[x_*;y_*]=[[x_*]_1;...;[x_*]_K;[y_*]_1;...;[y_*]_K]$ is an optimal solution to the convex optimization problem

$$\mathrm{Opt}=\max_{x\in X,y\in Y}\left[\sum_{k=1}^K\psi_k(x_k,y_k)\right],\tag{14}$$

and let

$$\phi_*^k(\omega_k)=\begin{cases}\xi_k^T\omega_k-\alpha_k,\ \xi_k=\frac{1}{2}\Sigma_k^{-1}[[x_*]_k-[y_*]_k],\\ \qquad\alpha_k=\frac{1}{2}\xi_k^T[[x_*]_k+[y_*]_k] & \text{for Gaussian } \mathcal{O}_k,\\ \frac{1}{2}\omega_k\ln\left([x_*]_k/[y_*]_k\right)-\frac{1}{2}[[x_*]_k-[y_*]_k] & \text{for scalar Poisson } \mathcal{O}_k,\\ \frac{1}{2}\ln\left([x_*]_{\omega_k}/[y_*]_{\omega_k}\right) & \text{for Discrete } \mathcal{O}_k.\end{cases}\tag{15}$$

Theorem 2.1 in our current situation implies the following statement:

Proposition 2.2 *In the framework described in section 2.1, assume that the observation scheme $\mathcal{O}^K$ is the direct product of some Gaussian, Discrete and scalar Poisson factors. Let $[x_*; y_*]$ be an optimal solution to the convex optimization problem (14) associated via the above construction with $\mathcal{O}^K$, and let*

$$\varepsilon_\star = \exp\{\mathrm{Opt}/2\}.$$

Then the error probabilities of the simple test with detector $\phi_*^a(\omega^K) = \sum_{k=1}^K \phi_*^k(\omega_k) - a$, where $\phi_*^k(\cdot)$ are as in (15), and $a \in \mathbb{R}$, satisfy

$$\epsilon_X \le \exp\{a\}\varepsilon_\star, \quad \textit{and} \quad \epsilon_Y \le \exp\{-a\}\varepsilon_\star.$$

Besides this, no test can distinguish between these hypotheses with the risk of test less than $\varepsilon_\star^2/4$.

Remarks. Two important remarks are in order.

When $\mathcal{O}^K$ is a direct product of Gaussian, Poisson and Discrete factors, finding the near-optimal simple test reduces to solving explicit well-structured convex optimization problem *with sizes polynomial in K and the maximal dimensions m_k of the factors*, and thus can be done in reasonable time, whenever K and $\max_k m_k$ are "reasonable." This is so in spite of the fact that the "formal sizes" of the saddle point problem associated with Φ could be huge (e.g., when all the factors $\mathcal{O}_k$ are discrete, the cardinality of Ω^K can grow exponentially with K, rapidly making a straightforward computation of Φ based on (2) impossible).

We refer to the indexes k and k', $1 \le k, k' \le K$, as equivalent in the direct product setup, augmented by convex compact subsets X, Y of $\mathcal{M}^K$, if $\mathcal{O}_k = \mathcal{O}_{k'}$, $x_k = x_{k'}$ for all $x \in X$, and $y_k = y_{k'}$ for all $y \in Y$. Denoting by K' the number of equivalence classes of indexes, it is clear that problem (14) is equivalent to a problem of completely similar structure, but with K' in the role of K. It follows that *the complexity of solving (14) is not affected by how large is the number K of factors; what matters is the number K' of equivalence classes of the indexes.* Similar phenomenon takes place when X and Y are direct products of their projections, X_k and Y_k, on the factors $\mathcal{M}_k$ of $\mathcal{M}^K$, and the equivalence of indexes k, k' is defined as $\mathcal{O}_k = \mathcal{O}_{k'}$, $X_k = X_{k'}$, $Y_k = Y_{k'}$.

3 Multiple hypotheses case

The examples outlined in section 2.3 demonstrate that the efficiently computable "nearly optimal" simple testing of composite hypotheses suggested by Theorem 2.1 and Proposition 2.2, while imposing strong restrictions on the underlying observation scheme, covers nevertheless some interesting and important applications. This testing "as it is," however, deals only with "dichotomies" (pairs of hypotheses) of special structure. In this section, we intend to apply our results to the situation when we should decide on more than two hypotheses, or still on two hypotheses, but more complicated than those considered in Theorem 2.1. Our general setup here is as follows. We are given a Polish observation space Ω along with a collection $X_1, ..., X_M$ of (nonempty) families of Borel probability distributions on Ω. Given an observation ω drawn from a distribution p *belonging to the union of these families* (pay attention to this default assumption!), we want to make some conclusions on the "location" of p. We will be interested in questions of two types:

A. [testing multiple hypotheses] We want to identify the family (or families) in the collection to which p belongs.

B. [testing unions] Assume our families $X_1, ..., X_M$ are split into two groups – "red" and "blue" families. The question is, whether p belongs to a red or a blue family.

When dealing with these questions, we will assume that for some pairs (i, j), $i \neq j$, of indexes from $1, ..., M$ (let the set of these pairs be denoted $\mathcal{I}$) we are given "pairwise tests" T_{ij} deciding on the pairs of hypotheses H_i, H_j (where H_k states that $p \in X_k$). To avoid ambiguities, we assume once for ever that the only possible outcomes of a test T_{ij} are either to reject H_i (and accept H_j), or to reject H_j (and accept H_j). For $(i, j) \in \mathcal{I}$, we are given the risks ϵ_{ij} (an upper bound on the probability for T_{ij} to reject H_i when $p \in X_i$) and $\bar{\epsilon}_{ij}$ (an upper bound on the probability for T_{ij} to reject H_j when $p \in X_j$). We suppose that whenever $(i, j) \in \mathcal{I}$, so is (j, i), and the tests T_{ij} and T_{ji} are the same, meaning that when run on an observation ω, T_{ij} accepts H_i if and only if T_{ji} accepts H_i. In this case we lose nothing when assuming that $\epsilon_{ij} = \bar{\epsilon}_{ji}$.

Our goal in this section is to "assemble" the pairwise tests T_{ij} into a test for deciding on "complex" hypotheses mentioned in A and in B. For example, assuming that T_{ij}'s are given for all pairs i, j with $i \neq j$, the simplest test for A would be as follows: given observation ω, we run on it tests T_{ij} for every pair i, j with $i \neq j$, and accept H_i when all tests T_{ij} with $j \neq i$ accept H_i. As a result of this procedure, at most one of the hypotheses will be accepted. Applying the union bound, it is immediately seen that if ω is drawn from p belonging to some X_i, H_i will be rejected with probability at most $\sum_{j \neq i} \epsilon_{ij}$, so that the quantity $\max_i \sum_{j \neq i} \epsilon_{ij}$ can be considered as the risk of our aggregated test.

The point in what follows is that when T_{ij} are tests of the type yielded by Theorem 2.1, we have wider "assembling options". Specifically, we will consider the case where

- T_{ij} are "simple tests induced by detectors ϕ_{ij}," where $\phi_{ij}(\omega) : \Omega \to \mathbb{R}$ are Borel functions; given ω, T_{ij} accepts H_i when $\phi_{ij}(\omega) > 0$, and accepts H_j when $\phi_{ij}(\omega) < 0$, with somehow resolved "ties" $\phi_{ij}(\omega) = 0$. To make T_{ij} and T_{ji} "the same," we will always assume that

$$\phi_{ij}(\omega) \equiv -\phi_{ji}(\omega), \ \omega \in \Omega, \ (i, j) \in \mathcal{I}. \tag{16}$$

- The risk bounds ϵ_{ij} "have a specific origin", namely, they are such that for all $(i, j) \in \mathcal{I}$,

$$(a) \quad \int_\Omega \exp\{-\phi_{ij}(\omega)\} p(d\omega) \leq \epsilon_{ij} \ \forall p \in X_i; \quad (b) \quad \int_\Omega \exp\{\phi_{ij}(\omega)\} p(d\omega) \leq \bar{\epsilon}_{ij}, \ \forall p \in X_j. \tag{17}$$

In the sequel, we refer to the quantities $\widehat{\epsilon}_{ij} := \sqrt{\epsilon_{ij}\bar{\epsilon}_{ij}}$ as to the *risks* of the detectors ϕ_{ij}. Note that the simple tests provided by Theorem 2.1 meet the just outlined assumptions. Another example is the one where X_i are singletons, and the distribution from X_i has density $p_i(\cdot) > 0$ with respect to a common for all i measure P on Ω; setting $\phi_{ij}(\cdot) = \frac{1}{2} \ln(p_i(\cdot)/p_j(\cdot))$ (so that T_{ij} are the standard likelihood ratio tests) and specifying $\epsilon_{ij} = \bar{\epsilon}_{ij}$ as Hellinger affinities of p_i and p_j, we meet our assumptions. Furthermore, *every* collection of pairwise tests $\overline{T}_{ij}$, $(i, j) \in \mathcal{I}$, deciding, with risks $\delta_{ij} = \delta_{ji} \in (0, 1/2)$, on the hypotheses H_i, H_j, $(i, j) \in \mathcal{I}$, gives rise to pairwise detectors ϕ_{ij} meeting (16) and (17) with $\epsilon_{ij} = \bar{\epsilon}_{ij} = 2\sqrt{\delta_{ij}(1 - \delta_{ij})}$ (cf. remark after Theorem 2.1). Indeed, to this end it suffices to set $\phi_{ij}(\omega) = \frac{1}{2} \ln\left(\frac{1-\delta_{ij}}{\delta_{ij}}\right) \overline{T}_{ij}(\omega)$ where, clearly, $\overline{T}_{ij}(\omega) = -\overline{T}_{ji}(\omega)$.

The importance of the above assumptions becomes clear from the following immediate observations:

1. By evident reasons, (17.a) and (17.b) indeed imply that when $(i,j) \in \mathcal{I}$ and $p \in X_i$, the probability for T_{ij} to reject H_i is $\leq \epsilon_{ij}$, while when $p \in X_j$, the probability for the test to reject H_j is $\leq \bar{\epsilon}_{ij}$. Besides this, taking into account that $\phi_{ij} = -\phi_{ji}$, we indeed ensure $\epsilon_{ij} = \bar{\epsilon}_{ji}$;

2. Relations (17.a) and (17.b) are preserved by a shift of the detector – by passing from $\phi_{ij}(\cdot)$ to $\phi_{ij}(\cdot) - a$ (accompanied with passing from ϕ_{ji} to $\phi_{ji} + a$) and simultaneous passing from ϵ_{ij}, $\bar{\epsilon}_{ij}$ to $\exp\{a\}\epsilon_{ij}$ and $\exp\{-a\}\bar{\epsilon}_{ij}$. In other words, all what matters is the product $\epsilon_{ij}\bar{\epsilon}_{ij}$ (i.e., the squared risk $\widehat{\epsilon}^2_{ij}$ of the detector ϕ_{ij}), and we can "distribute" this product between the factors as we wish, for example, making $\epsilon_{ij} = \bar{\epsilon}_{ij} = \widehat{\epsilon}_{ij}$;

3. Our assumptions are "ideally suited" for passing from a single observation ω drawn from a distribution $p \in \bigcup_{i=1}^{M} X_i$ to observing a K-tuple $\omega^K = (\omega_1, ..., \omega_K)$ of observations drawn, independently of each other, from p. Indeed, setting $\phi^K_{ij}(\omega_1, ..., \omega_K) = \sum_{k=1}^{K} \phi_{ij}(\omega_k)$, relations (17.a) and (17.b) clearly imply similar relations for ϕ^K_{ij} in the role of ϕ_{ij} and $[\epsilon_{ij}]^K$ and $[\bar{\epsilon}_{ij}]^K$ in the role of ϵ_{ij} and $\bar{\epsilon}_{ij}$. In particular, when $\max(\epsilon_{ij}, \bar{\epsilon}_{ij}) < 1$, passing from a single observation to K of them rapidly decreases the risks as K grows.

4. The left hand sides in relations (17.a) and (17.b) are linear in p, so that (17) remains valid when the families of probability distributions X_i are extended to their convex hulls.

In the rest of this section, we derive "nontrivial assemblings" of pairwise tests, meeting the just outlined assumptions, in the context of problems A and B.

3.1 Testing unions

3.1.1 Single observation case

Let us assume that we are given a family $\mathcal{P}$ of probability measures on a Polish space Ω equipped with a σ-additive σ-finite Borel measure P, and all distributions from $\mathcal{P}$ have densities w.r.t. P; we identify the distributions from $\mathcal{P}$ with these densities. Let $X_i \subset \mathcal{P}$, $i = 1, ..., m$ and $Y_j \subset \mathcal{P}$, $j = 1, ..., n$. Assume that pairwise detectors – Borel functions $\phi_{ij}(\cdot) : \Omega \to \mathbb{R}$, with risk bounded with $\epsilon_{ij} > 0$, are available for all pairs (X_i, Y_j), $i = 1, ..., m$, $j = 1, ..., n$, namely,

$$\int_\Omega \exp\{-\phi_{ij}(\omega)\}p(\omega)P(d\omega) \leq \epsilon_{ij}, \ \forall p \in X_i, \qquad \int_\Omega \exp\{\phi_{ij}(\omega)\}q(\omega)P(d\omega) \leq \epsilon_{ij}, \ \forall q \in Y_j.$$

Consider now the problem of deciding between the hypotheses

$$H_X : \ p \in X = \bigcup_{i=1}^{m} X_i \ \text{ and } \ H_Y : \ p \in Y = \bigcup_{j=1}^{n} Y_j.$$

on the distribution p of observation ω.

Let $E = [\epsilon_{ij}]_{i,j} \in \mathbb{R}^{m \times n}$. Consider the matrix $H = \begin{bmatrix} & E \\ E^T & \end{bmatrix}$. This is a symmetric entrywise nonzero nonnegative matrix. Invoking the Perron-Frobenius theorem, the leading eigenvalue of this matrix (which is nothing but the spectral norm $\|E\|_{2,2}$ of E) is positive, and

the corresponding eigenvector can be selected to be nonnegative. Let us denote this vector $z = [g; h]$ with $g \in \mathbb{R}^m_+$ and $h \in \mathbb{R}^n_+$, so that

$$Eh = \|E\|_{2,2} g, \ E^T g = \|E\|_{2,2} h. \tag{18}$$

We see that if one of the vectors g, h, is zero, both are so, which is impossible. Thus, both g and h are nonzero nonnegative vectors; since E has all entries positive, (18) says that in fact g and h are positive. Therefore we can set

$$\begin{array}{rcl} a_{ij} & = & \ln(h_j/g_i), \ 1 \le i \le m, 1 \le j \le n, \\ \phi(\omega) & = & \max\limits_{i=1,...,m} \min\limits_{j=1,...,n} [\phi_{ij}(\omega) - a_{ij}] : \Omega \to \mathbb{R}. \end{array} \tag{19}$$

Given observation ω, we accept H_X when $\phi(\omega^K) \ge 0$, and accept H_Y otherwise.

Proposition 3.1 *In the described situation, we have*

$$\begin{array}{ll} (a) & \int_\Omega \exp\{-\phi(\omega)\} p(\omega) P(d\omega) \le \varepsilon := \|E\|_{2,2}, \ p \in X, \\ (b) & \int_\Omega \exp\{\phi(\omega)\} p(\omega) P(d\omega) \le \varepsilon, \ p \in Y. \end{array} \tag{20}$$

As a result, the risk of the just described test when testing H_X versus H_Y does not exceed $\varepsilon = \|E\|_{2,2}$.

3.1.2 Case of repeated observations

The above construction and result admit immediate extension onto the case of non-stationary repeated observations. Specifically, consider the following situation. For $1 \le t \le K$, we are given

1. Polish space Ω_t equipped with Borel σ-additive σ-finite measure P_t,
2. A family $\mathcal{P}_t$ of Borel probability densities, taken w.r.t. P_t, on Ω_t,
3. Nonempty sets $X_{it} \subset \mathcal{P}_t$, $Y_{jt} \subset \mathcal{P}_t$, $i \in \mathcal{I}_t = \{1, ..., m_t\}$, $j \in \mathcal{J}_t = \{1, ..., n_t\}$,
4. *Detectors* – Borel functions $\phi_{ijt}(\cdot) : \Omega_t \to \mathbb{R}$, $i \in \mathcal{I}_t$, $j \in \mathcal{J}_t$, along with positive reals ϵ_{ijt}, $i \in \mathcal{I}_t$, $j \in \mathcal{J}_t$, such that

$$\begin{array}{ll} (a) & \int_{\Omega_t} \exp\{-\phi_{ijt}(\omega)\} p(\omega) P_t(d\omega) \le \epsilon_{ijt} \ \forall (i \in \mathcal{I}_t, j \in \mathcal{J}_t, p \in X_{it}), \\ (b) & \int_{\Omega_t} \exp\{\phi_{ijt}(\omega)\} p(\omega) P_t(d\omega) \le \epsilon_{ijt} \ \forall (i \in \mathcal{I}_t, j \in \mathcal{J}_t, p \in Y_{jt}), \end{array} \tag{21}$$

Given time horizon K, consider two hypotheses on observations $\omega^K = (\omega_1, ..., \omega_K)$, $\omega_t \in \Omega_t$, $H_1 := H_X$ and $H_2 := H_Y$, as follows. According to hypothesis H_χ, $\chi = 1, 2$, the observations ω_t, $t = 1, 2, ..., K$, are generated as follows:

> "In the nature" there exists a sequence of "latent" random variables $\zeta_{1,\chi}, \zeta_{2,\chi}, \zeta_{3,\chi}, ...$ such that ω_t, $t \le K$, is a deterministic function of $\zeta_\chi^t = (\zeta_{1,\chi}, ..., \zeta_{t,\chi})$, and the conditional, ζ_χ^{t-1} being fixed, distribution of ω_t has density $p_t \in \mathcal{P}_t$ w.r.t. P_t, the density p_t being a deterministic function of ζ_χ^{t-1}. Moreover, *when $\chi = 1$, p_t belongs to $X_t := \bigcup\limits_{i \in \mathcal{I}_t} X_{ti}$, and when $\chi = 2$, it belongs to $Y_t := \bigcup_{j \in \mathcal{J}_t} Y_{jt}$.*

Our goal is to decide from observations $\omega^K = (\omega_1, ..., \omega_K)$ on the hypotheses H_X and H_Y.

The test we intend to consider is as follows. We set

$$E_t = [\epsilon_{ijt}]_{i,j} \in \mathbb{R}^{m_t \times n_t},\ H_t = \left[\begin{array}{cc} & E_t \\ E_t^T & \end{array}\right] \in \mathbb{R}^{(m_t+n_t)\times(m_t+n_t)},\ \varepsilon_t = \|E_t\|_{2,2}. \tag{22}$$

As above, the leading eigenvalue of the symmetric matrix H_t is ε_t, the corresponding eigenvector $[g^t; h^t]$, $g^t \in \mathbb{R}^{m_t}, h^t \in \mathbb{R}^{n_t}$ can be selected to be positive, and we have

$$E_t h^t = \varepsilon_t g^t,\ E_t^T g^t = \varepsilon_t h^t. \tag{23}$$

We set

$$\begin{array}{rcl} a_{ijt} & = & \ln(h_j^t/g_i^t),\ 1 \le i \le m_t, 1 \le j \le n_t, \\ \phi_t(\omega_t) & = & \max\limits_{i=1,...,m_t} \min\limits_{j=1,...,n_t} [\phi_{ijt}(\omega_t) - a_{ijt}] : \Omega \to \mathbb{R}, \\ \phi^K(\omega^K) & = & \sum_{t=1}^K \phi_t(\omega_t). \end{array} \tag{24}$$

Given observation $\omega^K = (\omega_1, ..., \omega_K)$, we accept H_X when $\phi^K(\omega^K) \ge 0$, and accept H_Y otherwise.

We have the following analogue of Proposition 2.2

Proposition 3.2 *In the situation of this section, we have*

$$\begin{array}{ll} (a) & \int_\Omega \exp\{-\phi_t(\omega)\} p(\omega) P(d\omega) \le \varepsilon_t := \|E_t\|_{2,2},\ p \in X_t, t = 1, 2, ... \\ (b) & \int_\Omega \exp\{\phi_t(\omega)\} p(\omega) P(d\omega) \le \varepsilon_t,\ p \in Y_t, t = 1, 2, ... \end{array} \tag{25}$$

As a result, the risk of the just described test does not exceed $\prod_{t=1}^K \varepsilon_t$.

Some remarks are in order.

Symmeterizing the construction. Inspecting the proof of Proposition 3.2, we see that the validity of its risk-related conclusion is readily given by the validity of (25). The latter relation, in turn, is ensured by the described in (24) scheme of "assembling" the detectors $\phi_{ijt}(\cdot)$ into $\phi_t(\cdot)$, but this is not the only assembling ensuring (25). For example, swapping X_t and Y_t, applying the assembling (24) to these "swapped" data and "translating" the result back to the original data, we arrive at the detectors

$$\overline{\phi}_t(\omega) = \min_{j=1,...,n_t} \max_{i=1,...,m_t} [\phi_{ijt}(\omega) - a_{ijt}],$$

with a_{ijt} given by (24), and these new detectors, when used in the role of ϕ_t, still ensure (25). Denoting by $\underline{\phi}_t$ the detector ϕ_t given by (24), observe that $\underline{\phi}_t(\cdot) \le \overline{\phi}_t(\cdot)$, and this inequality in general is strict. Inspecting the proof of Proposition 3.2, it is immediately seen that Proposition remains true whenever $\phi^K(\omega^K) = \sum_{t=1}^K \phi_t(\omega_t)$ with $\phi_t(\cdot)$ satisfying the relations

$$\underline{\phi}_t(\cdot) \le \phi_t(\cdot) \le \overline{\phi}_t(\cdot),$$

for example, with the intrinsically symmetric "saddle point" detectors

$$\phi_t(\cdot) = \max_{\lambda \in \Delta_{m_t}} \min_{\mu \in \Delta_{n_t}} \sum_{i,j} \lambda_i \mu_j [\phi_{ijt}(\cdot) - a_{ijt}] \qquad [\Delta_k = \{x \in \mathbb{R}^k : x \ge 0, \textstyle\sum_{i=1}^k x_i = 1\}]$$

Needless to say, similar remarks hold true in the context of Proposition 3.1, which is nothing but the stationary (i.e., with $K = 1$) case of Proposition 3.2.

Testing convex hulls. As it was already mentioned, the risk-related conclusions in Propositions 3.1, 3.2 depend solely on the validity of relations (20), (25). Now, density $p(\cdot)$ enters the left hand sides in (20), (25) linearly, implying that when, say, (25) holds true for some X_t, Y_t, the same relation holds true when the families of probability densities X_t, Y_t are extended to their convex hulls. Thus, in the context of Propositions 3.1, 3.2 we, instead of speaking about testing *unions*, could speak about testing *convex hulls* of these unions.

Simple illustration. Let p be a positive probability density on the real axis $\Omega = \mathbb{R}$ such that setting $\rho_i = \int \sqrt{p(\omega)p(\omega - i)}d\omega$, we have $\varepsilon := 2\sum_{i=1}^{\infty} \rho_i < \infty$. Let $p_i(\omega) = p(\omega - i)$, and let $I = \{\imath_1 < ... < \imath_m\}$ and $J = \{\jmath_1 < ... < \jmath_n\}$ be two non-overlapping finite subsets of $\mathbb{Z}$. Consider the case where $X_{it} = \{p_{\imath_i}(\cdot)\}$, $1 \le i \le m = m_t$, $Y_{jt} = \{p_{\jmath_j}(\cdot)\}$, $1 \le j \le n = n_t$, are singletons, and let us set

$$\begin{array}{rcl} \phi_{ijt}(\omega) & = & \frac{1}{2}\ln(p_{\imath_i}(\omega)/p_{\jmath_j}(\omega)),\ 1 \le i \le m,\ 1 \le j \le n, \\ \epsilon_{ijt} & = & \int \sqrt{p_{\imath_i}(\omega)/p_{\jmath_j}(\omega)}d\omega,\ 1 \le i \le m,\ 1 \le j \le n. \end{array}$$

This choice clearly ensures (21), and for the associated matrix $E_t \equiv E$ we have $\|E\|_{2,2} \le \varepsilon$.[6] Thus, when ε is small, we can decide with low risk on the hypotheses associated with $X_t := \bigcup_{i=1}^{m} X_{it}$, $Y_t := \bigcup_{j=1}^{n} Y_{jt}$; note that ε is independent of the magnitudes of m, n. Moreover, when $\varepsilon < 1$, and repeated observations, of the structure considered in Proposition 3.2, are allowed, $K = \lfloor \ln(1/\epsilon)/\ln(1/\varepsilon) \rfloor$ observations are sufficient to get a test with risk $\le \epsilon$, and K again is not affected by the magnitudes of m, n. Finally, invoking the above remark, we can replace in these conclusions the finite sets of probability densities X_t, Y_t with their convex hulls.

3.2 Testing multiple hypotheses

Let $X_1, ..., X_m$ be nonempty sets in the space of Borel probability distributions on a Polish space Ω, $E = [\epsilon_{ij}]$ be a symmetric $m \times m$ matrix with zero diagonal and positive off-diagonal entries, and let

$$\phi_{ij}(\omega) = -\phi_{ji}(\omega) : \Omega \to \mathbb{R},\ 1 \le i, j \le m,\ i \ne j,$$

be Borel *detectors* such that

$$\forall (i, j, 1 \le i, j \le m, i \ne j) : \int_{\Omega} \exp\{-\phi_{ij}(\omega)\}p(d\omega) \le \epsilon_{ij}\ \forall p \in X_i. \tag{26}$$

Given a skew-symmetric matrix $[\alpha_{ij}]_{1 \le i,j \le m}$ and setting $\bar{\phi}_{ij}(\cdot) = \phi_{ij}(\cdot) - \alpha_{ij}$, we get

$$\forall (i, j, 1 \le i, j \le m, i \ne j) : \int_{\Omega} \exp\{-\bar{\phi}_{ij}(\omega)\}p(d\omega) \le \exp\{\alpha_{ij}\}\epsilon_{ij}\ \forall p \in X_i. \tag{27}$$

Consider the following test aimed to decide, given an observation ω drawn from a distribution p known to belong to $X = \bigcup_{i=1}^{m} X_i$, on i such that $p \in X_i$ (we refer to the validity of the latter

[6] We use the following elementary fact: *Let E be a matrix with sums of magnitudes of entries in every row and every column not exceeding r. Then $\|E\|_{2,2} \le r$.* To be on the safe side, here is the proof: let $F = \begin{bmatrix} & E \\ E^T & \end{bmatrix}$, so that $\|E\|_{2,2} = \|F\|_{2,2}$, and $\|F\|_{2,2}$ is just the spectral radius of F. We clearly have $\|Fx\|_\infty \le r\|x\|_\infty$ for all x, whence the spectral radius of F is at most r.

inclusion as to hypothesis H_i). The test is as follows: we compute $\bar{\phi}_{ij}(\omega)$ for all $i \neq j$, and accept all H_i's such that all the quantities $\bar{\phi}_{ij}(\omega)$ with j distinct from i are positive. Note that since $\bar{\phi}_{ij}(\cdot) \equiv -\bar{\phi}_{ji}(\cdot)$, if some H_i is accepted by our test, no $H_{i'}$ with i' different from i can be accepted; thus, our test, for every ω, accepts at most one of the hypotheses H_i. Let us denote by ϵ_i the maximal, over $p \in X_i$, probability for the test to reject H_i when our observation ω is drawn from $p(\cdot)$. Note that since our test accepts at most one of H_i's, for every i the probability to accept H_i when the observation ω is drawn from a distribution $p(\cdot) \in X \backslash X_i$ (i.e., when H_i is false) does not exceed $\max_{j:j\neq i} \epsilon_j$.

Now recall that the risks ϵ_i depend on the shifts α_{ij}, and consider the problem as follows. Given "importance weights" $p_i > 0$, $1 \leq i \leq m$, we now aim to find the shifts α_{ij} resulting in the smallest possible quantity

$$\epsilon := \max_{1\leq i\leq m} p_i \epsilon_i,$$

or, more precisely, the smallest possible natural upper bound ε on this quantity. We define this bound as follows.

Let, for some i, an observation ω be drawn from a distribution $p \in X_i$. Given this observation, H_i will be rejected if for some $j \neq i$ the quantity $\bar{\phi}_{ij}(\omega)$ is nonpositive. By (26), for a given $j \neq i$, p-probability of the event in question is at most $\exp\{\alpha_{ij}\}\epsilon_{ij}$, which implies the upper bound on ϵ_i, specifically, the bound

$$\varepsilon_i = \sum_{j\neq i} \exp\{\alpha_{ij}\}\epsilon_{ij} = \sum_{j=1}^{m} \exp\{\alpha_{ij}\}\epsilon_{ij}$$

(recall that $\epsilon_{ii} = 0$ for all i). Thus, we arrive at the upper bound

$$\varepsilon := \max_i p_i \varepsilon_i = \max_i \sum_{j=1}^{m} p_i \epsilon_{ij} \exp\{\alpha_{ij}\} \tag{28}$$

on ϵ. What we want is to select $\alpha_{ij} = -\alpha_{ji}$ minimizing this bound.

Our goal is relatively easy to achieve: all we need is to solve the *convex* optimization problem

$$\varepsilon_* = \min_{\alpha=[\alpha_{ij}]} \left\{ f(\alpha) := \max_{1\leq i\leq m} \sum_j p_i \epsilon_{ij} \exp\{\alpha_{ij}\} : \alpha = -\alpha^T \right\}. \tag{29}$$

The problem (29) allows for a "closed form" solution.

Proposition 3.3 *Let ρ be the Perron-Frobenius eigenvalue of the entry-wise nonnegative matrix $\bar{E} = [p_i\epsilon_{ij}]_{1\leq i,j\leq m}$. The corresponding eigenvector $g \in \mathbb{R}^m$ can be selected to be* positive, *and for the choice $[\bar{\alpha}_{ij} := \ln(g_j) - \ln(g_i)]_{i,j}$, $1 \leq i, j \leq m$, one has $\varepsilon_* = f(\bar{\alpha}) = \rho$.*

Remark. The proof of Proposition 3.3 demonstrates that with the optimal assembling given by $\alpha_{ij} = \bar{\alpha}_{ij}$ all the quantities $p_i\epsilon_i$ in (28) become equal to $\varepsilon_* = \rho$. In particular, when $p_i = 1$ for all i, for every i the probabilities to reject H_i when the hypothesis is true, and to accept H_i when the hypothesis is false, are upper bounded by ρ.

3.2.1 A modification

In this section we focus on multiple hypothesis testing in the case when all importance factors p_i are equal to 1. Note that in this case the result we have just established can be void when the optimal value ε_* in (29) is ≥ 1, as this is the case, e.g., when some X_i and X_j with $i \neq j$ intersect. In the latter case, for every pair i, j with $i \neq j$ and $X_i \cap X_j \neq \emptyset$, the best – resulting in the smallest possible value of ϵ_{ij} – selection of ϕ_{ij} is $\phi_{ij} \equiv 0$, resulting in $\epsilon_{ij} = 1$. It follows that even with K-repeated observations (for which ϵ_{ij} should be replaced with ϵ_{ij}^K) the optimal value in (29) is ≥ 1, so that our aggregated test allows for only trivial bound $\varepsilon \leq 1$ on ε, see (28).[7] Coming back to the general situation where $p_i \equiv 1$ and ε_* is large, what can we do? A solution, applicable when $\epsilon_{ij} < 1$ for all i, j, is to pass to K-repeated observations; as we have already mentioned, this is equivalent to passing from the original matrix $E = [\epsilon_{ij}]$ to its entrywise power $E^{(K)} = [\epsilon_{ij}^K]$; when K is large, the leading eigenvalue ρ_K of $E^{(K)}$ becomes small. The question is what to do if some of ϵ_{ij} indeed are equal to 1, and a somewhat partial solution in this case may be obtained by substituting our original goal of highly reliable recovery of the true hypothesis with a less ambitious one. A natural course of action could be as follows. Let $\mathcal{I}$ be the set of all ordered pairs (i, j) with $1 \leq i, j \leq m$, and let $\mathcal{C}$ be a given subset of this set containing all "diagonal" pairs (i, i). We interpret the inclusion $(i, j) \in \mathcal{C}$ as the claim that H_j is "close" to H_i. [8] Imagine that what we care about when deciding on the collection of hypotheses $H_1, ..., H_m$ is not to miss a correct hypothesis and, at the same time, to reject all hypotheses which are "far" from the true one(s). This can be done by test as follows. Let us shift somehow the original detectors, that is, pass from $\phi_{ij}(\cdot)$ to the detectors $\phi'_{ij}(\cdot) = \phi_{ij}(\cdot) - \alpha_{ij}$ with $\alpha_{ij} = -\alpha_{ij}$, thus ensuring that

$$\phi'_{ij}(\cdot) := -\phi'_{ji}(\cdot) \ \& \ \int_\Omega \exp\{-\phi'_{ij}(\omega)\} p(d\omega) \leq \epsilon'_{ij} := \exp\{\alpha_{ij}\}\epsilon_{ij} \ \forall p \in X_i. \tag{30}$$

Consider the test as follows:

> **Test $\mathcal{T}$:** Given observation ω, we compute the matrix $[\phi'_{ij}(\omega)]_{ij}$. Looking one by one at the rows $i = 1, 2, ...m$ of this matrix, we accept H_i if all the entries $\phi'_{ij}(\omega)$ with $(i, j) \notin \mathcal{C}$ are positive, otherwise we reject H_i.
>
> The outcome of the test is the collection of all accepted hypotheses (which now is not necessary either empty or a singleton).

What we can say about this test is the following. Let

$$\epsilon = \max_i \sum_{j:(i,j)\notin\mathcal{C}} \epsilon'_{ij}, \tag{31}$$

and let the observation ω the test is applied to be drawn from distribution $p \in X_{i_*}$, for some i_*. Then

- if, for some $i \neq j$, $\mathcal{T}$ accepts both H_i and H_j, then either H_j is close to H_i, or H_i is close to H_j, or both.

[7] Of course, the case in question is intrinsically difficult – here no test whatsoever can make all the risks ϵ_i less than 1/2.

[8] Here the set of ordered pairs $\mathcal{C}$ is not assumed to be invariant w.r.t. swapping the components of a pair, so that in general "H_j is close to H_i" is not the same as "H_i is close to H_j."

Indeed, if neither H_i is close to H_j, nor H_j is close to H_i, both H_i, H_j can be accepted only when $\phi'_{ij}(\omega) > 0$ and $\phi'_{ji}(\omega) > 0$, which is impossible due to $\phi'_{ij}(\cdot) = -\phi'_{ji}(\cdot)$.

- p-probability for the true hypothesis H_{i_*} not to be accepted is at most ϵ.
  Indeed, by (30), the p-probability for ϕ'_{i_*j} to be nonpositive does not exceed ϵ'_{i_*j}. With this in mind, taking into account the description of our test and applying the union bound, p-probability to reject H_{i_*} does not exceed $\sum_{j:(i_*,j)\notin\mathcal{C}} \epsilon'_{i_*j} \le \epsilon$.

- p-probability of the event $\mathcal{E}$ which reads *"at least one of the accepted H_i's is such that both $(i, i_*) \notin \mathcal{C}$ and $(i_*, i) \notin \mathcal{C}$"* (that is, neither i_* is close to i, nor i is close to i_*) does not exceed ϵ.
  Indeed, let I be the set of all those i for which $(i, i_*) \notin \mathcal{C}$ and $(i_*, i) \notin \mathcal{C}$. For a given $i \in I$, H_i can be accepted by our test only when $\phi'_{ii_*}(\omega) > 0$ (since $(i, i_*) \notin \mathcal{C}$), implying that $\phi'_{i_*i}(\omega) < 0$. By (30), the latter can happen with p-probability at most ϵ'_{i_*i}. Applying the union bound, the p-probability of the event $\mathcal{E}$ is at most

$$\sum_{i\in I} \epsilon'_{i_*i} \le \sum_{i:(i_*,i)\notin\mathcal{C}} \epsilon'_{i_*i} \le \epsilon$$

  (we have taken into account that whenever $i \in I$, we have $(i_*, i) \notin \mathcal{C}$, that is, $I \subset \{i : (i_*, i) \notin \mathcal{C}\}$).

When ϵ is small (which, depending on how closeness is specified, can happen even when some of ϵ'_{ij} are not small), the simple result we have just established is "better than nothing:" it says that up to an event of probability 2ϵ, the true hypotheses H_{i_*} is accepted, and all accepted hypotheses H_j are such that either j is close to i_*, or i_* is close to j, or both.

Clearly, given $\mathcal{C}$, we would like to select α_{ij} to make ϵ as small as possible. The punch line is that this task is relatively easy: all we need is to solve the *convex* optimization problem

$$\min_{[\alpha_{ij}]_{i,j}} \left\{ \max_{1\le i\le m} \sum_{j:(i,j)\notin\mathcal{C}} \epsilon_{ij} \exp\{\alpha_{ij}\} : \alpha_{ij} \equiv -\alpha_{ji} \right\}. \tag{32}$$

Special case: testing multiple unions. Consider the case when "closeness of hypotheses" is defined as follows: the set $\{1, ..., M\}$ of hypotheses' indexes is split into $L \ge 2$ nonempty non-overlapping subsets $\mathcal{I}_1, ..., \mathcal{I}_L$, and H_j is close to H_i if and only if both i, j belong to the same element of this partition. Setting $E = [\epsilon_{ij}]_{i,j}$, let $D = [\delta_{ij}]$ be the matrix obtained from E by zeroing out all entries ij with i, j belonging to $\mathcal{I}_\ell$ for some $1 \le \ell \le L$. Problem (32) now reads

$$\min_{[\alpha_{ij}]} \left\{ \max_{1\le i\le M} \sum_{1\le j\le M} \delta_{ij} \exp\{\alpha_{ij}\} : \alpha = -\alpha^T \right\}.$$

This problem, similarly to problem (29), admits a closed form solution: the Perron-Frobenius eigenvector g of the entrywise nonnegative symmetric matrix D can be selected to be positive, an optimal solution is given by $\alpha_{ij} = \ln(g_j) - \ln(g_i)$, and the optimal value is $\epsilon_* := \|D\|_{2,2}$. Test $\mathcal{T}$ associated with the optimal solution can be converted into a test $\widehat{\mathcal{T}}$ deciding on L hypotheses $\mathcal{H}_\ell = \bigcup_{i\in\mathcal{I}_\ell} H_i$, $1 \le \ell \le k$; specifically, when $\mathcal{T}$ accepts some hypothesis H_i, $\widehat{\mathcal{T}}$ accepts hypothesis

$\mathcal{H}_\ell$ with ℓ uniquely defined by the requirement $i \in \mathcal{I}_\ell$. The above results on $\mathcal{T}$ translate in the following facts about $\widehat{\mathcal{T}}$:

- $\widehat{\mathcal{T}}$ never accepts more than one hypothesis;
- let the observation ω on which $\widehat{\mathcal{T}}$ is run be drawn from a distribution p obeying, for some $1 \le i \le M$, the hypothesis H_i, and let ℓ be such that $i \in \mathcal{I}_\ell$. Then the p-probability for $\widehat{\mathcal{T}}$ to reject the hypothesis $\mathcal{H}_\ell$ is at most ϵ_*.

When $L = 2$ we come back to the situation considered in section 3.1.1, and what has just been said about $\widehat{\mathcal{T}}$ recovers the risk-related result of Proposition 3.1; moreover, when $L = 2$, the test $\widehat{\mathcal{T}}$ is, essentially, the test based on the detector ϕ given by (19).[9] Note that when $L > 2$, one could use the detector-based tests, yielded by the construction in section 3.1.1, to build "good" detectors for the pairs of hypotheses $\mathcal{H}_\ell$, $\mathcal{H}_{\ell'}$ and then assemble these detectors, as explained in section 3.2, into a test deciding on multiple hypotheses $\mathcal{H}_1, ..., \mathcal{H}_L$, thus getting an "alternative" to $\widehat{\mathcal{T}}$ test $\widetilde{\mathcal{T}}$. Though both tests are obtained by aggregating detectors ϕ_{ij}, $1 \le i, j \le M$, in the test $\widehat{\mathcal{T}}$ we aggregate them "directly", while the aggregation in test $\widetilde{\mathcal{T}}$ is done in two stages where we first assemble ϕ_{ij} into pairwise detectors $\widetilde{\phi}_{\ell\ell'}$ for $\mathcal{H}_\ell$, $\mathcal{H}_{\ell'}$, and then assemble these new detectors into a test for multiple hypotheses $\mathcal{H}_1, ..., \mathcal{H}_L$. However, the performance guarantees for the test $\widetilde{\mathcal{T}}$ can be only worse than those for the test $\widehat{\mathcal{T}}$ – informally, when assembling ϕ_{ij} into $\widetilde{\phi}_{\ell,\ell'}$, we take into account solely the "atomic contents" of the aggregated hypotheses $\mathcal{H}_\ell$ and $\mathcal{H}_{\ell'}$, that is, look only at the "atoms" H_i with $i \in \mathcal{I}_\ell \cup \mathcal{I}_{\ell'}$, while when assembling ϕ_{ij} into $\widehat{\mathcal{T}}$, we look at all m atoms simultaneously.[10]

Near-optimality. Let the observation scheme underlying the just considered "multiple unions" situation be K-repeated version $\mathcal{O}^K$ of a good observation scheme $\mathcal{O} = ((\Omega, P), \{p_\mu(\cdot) : \mu \in \mathcal{M}\}, \mathcal{F})$, meaning that our observation is $\omega = \omega^K := (\omega_1, ..., \omega_K)$ with ω_t drawn, independently of each other, from a distribution p, and i-th of our M hypotheses, H_i, states that p belongs to the set $X_i = \{p_\mu : \mu \in Q_i\}$, where Q_i are convex compact subsets of $\mathcal{M}$. Let ϕ_{ij} be the pairwise detectors for H_i and H_j yielded by Theorem 2.1, and let $\widehat{\mathcal{T}}^K$ be the test deciding on aggregated hypotheses $\mathcal{H}_\ell$'s from K-repeated observations ω^K and built by assembling detectors $\phi_{ij}^K = \sum_{t=1}^K \phi_{ij}(\omega_t)$. We have the following near-optimality result (cf. Proposition 2.1):

Proposition 3.4 *In the just described situation and given $\epsilon \in (0, 1/4)$, assume that in the nature there exists a test $\overline{T}$, based on $\bar{K}$-repeated observations $\omega^{\bar{K}}$, deciding on $\mathcal{H}_1, ..., \mathcal{H}_L$ and such that $\overline{T}$ never accepts more than one hypothesis and, for every $\ell \le L$, rejects $\mathcal{H}_\ell$ when the hypothesis is true with probability $\le \epsilon$. Then the same performance guarantees are shared by the test $\widehat{\mathcal{T}}^K$, provided that*

$$K \ge \frac{2\ln(M/\epsilon)}{\ln(1/\epsilon) - 2\ln 2}\bar{K}.$$

[9] The only subtle difference, completely unimportant in our context, is that the latter test accepts $\mathcal{H}_1$ whenever $\phi(\omega) \ge 0$ and accepts $\mathcal{H}_2$ otherwise, while $\widehat{\mathcal{T}}$ accepts $\mathcal{H}_1$ when $\phi(\omega) > 0$, accepts $\mathcal{H}_2$ when $\phi(\omega) < 0$ and accepts nothing when $\phi(\omega) = 0$.

[10] The formal reasoning is as follows. On a close inspection, to get risk bound $\widetilde{\epsilon}$ for $\widetilde{\mathcal{T}}$, we start with the $M \times M$ matrix D partitioned into $L \times L$ blocks $D^{\ell\ell'}$ (this partitioning is induced by splitting the indexes of rows and columns into the groups $\mathcal{I}_1$,...,$\mathcal{I}_L$), and form the $L \times L$ matrix G with entries $\gamma_{\ell\ell'} = \|D^{\ell\ell'}\|_{2,2}$; $\widetilde{\epsilon}$ is nothing but $\|G\|_{2,2}$, while the risk bound ϵ_* for $\widehat{\mathcal{T}}$ is $\|D\|_{2,2}$. Thus, $\epsilon_* \le \widetilde{\epsilon}$ by the construction of matrix G from D.

4 Case studies

4.1 Hypotheses testing in PET model

To illustrate applications of the simple test developed in section 2.3.3 we discuss here a toy testing problem in the *Positron Emission Tomography (PET) model*.

A model of PET which is accurate enough for medical purposes is as follows. The patient is injected a radioactive tracer and is placed inside a cylinder with the inner surface split into detector cells. Every tracer disintegration act gives rise to two γ-quants flying in opposite directions along a randomly oriented line (Line of Response, LOR) passing through the disintegration point. Unless the LOR makes too small angle with the cylinder's axis, the γ-quants activate (nearly) simultaneously a pair of detector cells; this event ("coincidence") is registered, and the data acquired in a PET study is the list of the detector pairs in which the coincidences occurred. The goal of the study is to infer about the density of the tracer on the basis of these observations.

After appropriate discretization of the field of view into small cells, disintegration acts in a particular cell form a Poisson processes with intensity proportional to the density of the tracer in the cell. The entries of the observations vector ω are indexed by *bins* i – pairs of detectors, ω_i being the number of coincidences registered during the study by bin i. Mathematically, ω_i, $i = 1, ..., m$, are the realizations of independent across i's Poisson random variables with parameters $\mu_i = (tP\lambda)_i$, where t is the observation time, λ is the vector of intensities of disintegration in the cells of the field of view, and the entries P_{ij} in the matrix P are the probabilities for a LOR originating in cell j to be registered by bin i; this matrix is readily given by the geometry of PET's device. We observe that PET model meets the specifications of what we call Poisson observation scheme.

Let $\mathcal{M}$ be the image, under the linear mapping $\lambda \mapsto tP\lambda$, of the set $\Lambda = \Lambda_{L,R}$ of non-vanishing on $\mathbb{R}^n$ densities λ satisfying some regularity restrictions, specifically, such that the uniform norm of discrete Laplacian of λ is upper-bounded by L, and the average of λ, over all pixels, is upper-bounded by R, i.e.

$$\Lambda_{L,R} = \left\{ \begin{array}{c} \lambda \in \mathbb{R}^n : \ \lambda \geq 0, \ n^{-1}\sum_{j=1}^n \lambda_j \leq R, \\ \frac{1}{4}|4\lambda_{j(k,\ell)} - \lambda_{j(k-1,\ell)} - \lambda_{j(k,\ell-1)} - \lambda_{j(k+1,\ell)} - \lambda_{j(k,\ell+1)}| \leq L, \ \ 1 \leq j \leq n \end{array} \right\},$$

(k, ℓ) being the coordinates of the cell j in the field of view (by convention, $\lambda_{j(k,\ell)} = 0$ when the cell (k, ℓ) is not in the field of view). Our goal is to distinguish two hypotheses, H_1 and H_2, about λ:

$$H_1 : \ \lambda \in \Lambda_1 = \{\lambda \in \Lambda : \ g(\lambda) \leq \alpha\}, \ \ H_1 : \ \lambda \in \Lambda_2 = \{\lambda \in \Lambda : \ g(\lambda) \geq \alpha + \rho\}, \qquad (\mathcal{P}_{g,\alpha}[\rho])$$

$g(\lambda) = g^T\lambda$ being a given linear functional of λ. From now on we assume that $g \notin \mathrm{Ker}(P)$ and $\rho > 0$, thus the described setting corresponds to the Poisson case of the hypotheses testing problem of section 2.3.3, $X = tP\Lambda_1$ and $Y = tP\Lambda_2$ being two nonintersecting convex sets of observation intensities. Let us fix the value $\epsilon \in (0, 1)$, and consider the optimization problem

$$t_* = \min_t \max_{\lambda,\lambda'} \left\{ t : \ \begin{array}{c} -\frac{t}{2}\sum_{i=1}^m \left[\sqrt{[P\lambda]_i} - \sqrt{[P\lambda']_i}\right]^2 \geq \ln \epsilon, \\ \lambda, \lambda' \in \Lambda, \ g(\lambda) \leq \alpha, \ g(\lambda') \geq \alpha + \rho. \end{array} \right\} \qquad (33)$$

Suppose that the problem parameters are such that both hypotheses in $(\mathcal{P}_{g,\alpha}[\rho])$ are not empty. It can be easily seen that in this case problem (33) is solvable and its optimal value t_* is

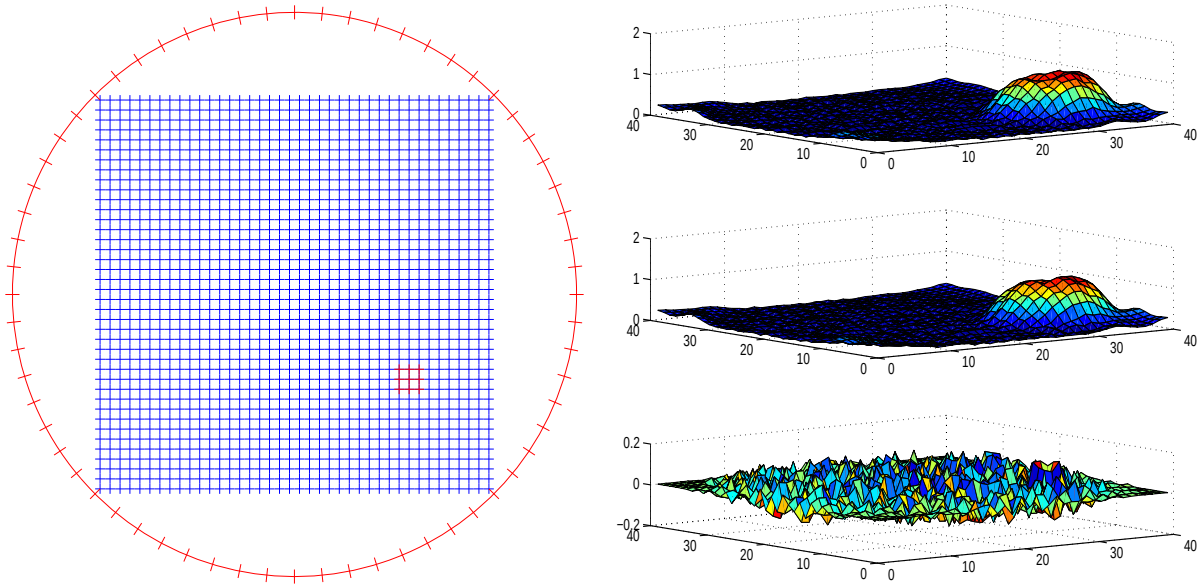

Figure 1: Toy PET experiment. Left: 40×40 field of view with 3×3 "suspicious spot" (in red) and the ring of 64 detector cells. Right: the hardest to distinguish tracer densities λ_* (top) and λ'_* (middle), and the difference of these densities (bottom).

positive 0. Let $[\lambda_*; \lambda'_*]$ be the $[\lambda; \lambda']$-component of an optimal solution to (33), consider the test T_* associated with the detector

$$\phi_*(\omega) = \tfrac{1}{2}\sum_{i=1}^{m} \ln\left[\frac{[P\lambda_*]_i}{[P\lambda'_*]_i}\right]\omega_i - \tfrac{1}{2}\sum_{i=1}^{m}[P\lambda_* - P\lambda'_*]_i. \tag{34}$$

By applying Theorem 2.1 in the Poisson case (cf. (12)) we conclude that the risk of the test T_* associated with detector ϕ_*, when applied to the problem testing problem ($\mathcal{P}_{g,\alpha}[\rho]$) is bounded with ϵ, as soon as the observation time $t \geq t_*$.

In the numerical experiment we are about to describe we simulate a 2D PET device with square field of view split into 40×40 pixels (i.e., dimension of λ was $n = 1600$). The detector cells are represented by $k = 64$ equal arcs of the circle circumscribing the field of view, resulting in the observation space (pairs of detectors which may be activated during the experiment) of dimension $m = 1536$. We choose $g(\cdot)$ to be the density average over a specific 3×3 "suspicious spot" (see the left plot on figure 1), and values of $\alpha = 1.0$ and $\rho = 0.1$, so that under H_1 the average of the density λ of the tracer on the spot is upper-bounded by 1, while under H_2 this average is at least 1.1. The regularity parameters of the density class $\Lambda_{L,R}$ were set to $L = 0.05$ and $R = 1$, the observation time t^* and parameters of the detector ϕ_* were selected according to (33) and (34) with $\epsilon = 0.01$.

On the right plot on figure 1 we present the result of computation of the hardest to distinguish densities $\lambda_* \in \Lambda_1$ and $\lambda'_* \in \Lambda_2$. We have also measured the actual performance of our test by simulating 2000 PET studies with varying from study to study density of the tracer. In the first 1000 of our simulations the true density was selected to obey H_1, and in the remaining

1000 simulations – to obey H_2, and we did our best to select the densities which make decision difficult. In the reported experiment the empirical probabilities to reject the true hypothesis were 0.005 when the true hypothesis was H_1, and 0.008 when the true hypothesis was H_2.

4.2 Event detection in sensor networks

4.2.1 Problem description

Suppose that m sensors are deployed on the domain $G \subseteq \mathbb{R}^d$. The signals are real-valued functions $x : \Gamma \to \mathbb{R}^n$ on a grid $\Gamma = (\gamma_i)_{i=1,...,n} \subset G$, and the observation ω_j delivered by jth sensor, $j = 1, ..., m$, is a linear form of the signal, contaminated with random noise. So we have at our disposal an observation $\omega \sim P_\mu$ – a random vector in $\mathbb{R}^m$ with the distribution parameterized by $\mu \in \mathbb{R}^m$, where $\mu = Ax$ and $A \in \mathbb{R}^{m\times n}$ is a known matrix of sensor responses (jth row of A is the response of the jth sensor). Further, we assume that the signal x can be decomposed into $x = s + v$, where $v \in \mathcal{V}$ is a background (nuisance) signal, $\mathcal{V}$ is a known convex and compact set in $\mathbb{R}^n$. We assume that at most one event can take place during the observation period, and an event occurring at a node γ_i of the grid produces the signal $s = re[i] \in \mathbb{R}^n$ on the grid of known signature $e[i]$ with unknown real factor r.

We want to decide whether an event occurred during the observation period, i.e. to test the null hypothesis that no event happened against the alternative that exactly one event took place. To make a consistent decision possible we need the alternative to be separated from the null hypothesis, so we require, first, that $Ae[i] \neq 0$ for all i, and, second, that under the alternative, when an event occurs at a node $\gamma_i \in \Gamma$, we have $s = re[i]$ with $|r| \geq \rho_i$ with some given $\rho_i > 0$. Thus we come to the testing problem as follows:

$$(\mathcal{D}_\rho) \qquad \begin{array}{l}\text{Given } \rho = [\rho_1; ...; \rho_n] > 0, \text{ test the hypothesis } H_0 : \ s = 0 \text{ against the} \\ \text{alternative } H_1(\rho) : s = re[i] \text{ for some } i \in \{1, ..., n\} \text{ and } r \text{ with } |r| \geq \rho_i.\end{array} \tag{35}$$

Our goal is, given an $\epsilon \in (0, 1)$, to construct a test with risk $\leq \epsilon$ for as wide as possible (i.e., with as small ρ as possible) alternative $H_1(\rho)$.

The problem of multi-sensor detection have recently received much attention in the signal processing and statistical literature (see e.g., [43, 44] and references therein). Furthermore, a number of classical detection problems, extensively studied in statistical literature, such as detecting jumps in derivatives of a function and cusp detection [2, 22, 23, 33, 39, 40, 45, 46], detecting a nontrivial signal on input of a dynamical system [25], or parameter change detection [4] can be posed as $(\mathcal{D}_\rho)$.

Our current objective is to apply the general approach described in section 3.1.1 to the problem $(\mathcal{D}_\rho)$. Note that, in terms of the parameter μ underlying the distribution of the observation ω, the hypothesis H_0 corresponds to $\mu \in X := A\mathcal{V}$, a convex compact set, while the alternative H_1 is represented by the union $Y = \bigcup_{i=1}^{n} Y_i$ of the sets $Y_i = \{Are[i] + \nu, \ \nu \in \mathcal{V}, |r| \geq \rho_i\}$. To comply with assumptions of section 2 we bound the sets Y_i by imposing an upper bound on the amplitude r of the useful signal: from now on we assume that $\rho_i \leq |r| \leq R$ in the definition of $(\mathcal{D}_\rho)$.[11]

[11] Imposing a finite upper bound R on $|r|$ is a minor (and non-restrictive, as far as applications are concerned) modification of the problem stated in the introduction; the purely technical reason for this modification is our desire to work with compact sets of parameters. It should be stressed that R does not affect the performance bounds to follow.

Given a test $\phi(\cdot)$ and $\epsilon > 0$, we call a collection $\rho = [\rho_1; ...; \rho_n]$ of positive reals *an ϵ-rate profile of the test ϕ* if whenever the signal s underlying our observation is $re[i]$ for some i and r with $\rho_i \le |r| \le R$, the hypothesis H_0 will be rejected by the test with probability $\ge 1-\epsilon$, whatever be the nuisance $v \in \mathcal{V}$, and whenever $s = 0$, the probability for the test to reject H_0 is $\le \epsilon$, whatever be the nuisance $v \in \mathcal{V}$. Our goal is to design a test with ϵ-rate profile "nearly best possible" in the sense of the following definition:

> Let $\kappa \ge 1$. A test T with risk ϵ in the problem $(\mathcal{D}_\rho)$ is said to be *κ–rate optimal*, if there is no test with the risk ϵ in the problem $(\mathcal{D}_{\underline{\rho}})$ with $\underline{\rho} < \kappa^{-1}\rho$ (inequalities between vectors are understood componentwise).

4.2.2 Poisson case

Let the sensing matrix A be nonnegative and without zero rows, let the signal x be nonnegative, and let the entries ω_i in our observation be independent and obeying Poisson distribution with the intensities $\mu := [\mu_1; ...; \mu_m] = Ax$. In this case the null hypothesis is that the signal is a pure nuisance:

$$H_0 : \ \mu \in X = \{\mu = Av, \ v \in \mathcal{V}\},$$

where $\mathcal{V}$ is the nuisance set assumed to be a nonempty compact convex set belonging to the interior of the nonnegative orthant. The alternative $H_1(\rho)$ is the union over $i = 1, ..., n$ of the hypotheses

$$H^i(\rho_i) : \ \mu \in Y(\rho_i) = \{rAe[i] + Av, \ v \in \mathcal{V}, \ \rho_i \le r \le R\},$$

where $e[i] \ge 0$, $1 \le i \le n$, satisfy $Ae[i] \ne 0$. For $1 \le i \le n$, let us set (cf. section 2.3.3)

$$\rho_i^P(\epsilon) = \max_{\rho,r,u,v} \left\{ \rho : \begin{array}{l} \frac{1}{2}\sum_{\ell=1}^m \left[\sqrt{[Au]_\ell} - \sqrt{[A(re[i]+v)]_\ell}\right]^2 \le \ln(\sqrt{n}/\epsilon) \\ u \in \mathcal{V}, v \in \mathcal{V}, r \ge \rho \end{array} \right\}, \qquad (P_\epsilon^i)$$

$$\phi_i(\omega) = \sum_{\ell=1}^m \ln(\sqrt{[Au^i]_\ell/[A(r^i e[i] + v^i)]_\ell})\omega_\ell - \frac{1}{2}\sum_{\ell=1}^m [A(u^i - r^i e[i] - v^i)]_\ell, \qquad (36)$$

where r^i, u^i, v^i are the r, u, v-components of an optimal solution to (P_ϵ^i) (of course, in fact $r^i = \rho_i^P(\epsilon)$). Finally, let

$$\rho^P[\epsilon] = [\rho_1^P(\epsilon); ...; \rho_n^P(\epsilon)], \quad \widehat{\phi}_P(\omega) = \min_{i=1,...,n} \phi_i(\omega) + \tfrac{1}{2}\ln(n).$$

Detector $\widehat{\phi}_P(\cdot)$ specifies a test which accepts H_0, the observation being ω, when $\widehat{\phi}_P(\omega) \ge 0$ (i.e., with observation ω, all pairwise tests with detectors ϕ_i, $1 \le i \le n$, $\chi = \pm 1$, when deciding on H_0 vs. H^i, accept H_0), and accepts $H_1(\rho)$ otherwise.

Proposition 4.1 *Whenever $\rho \ge \rho^P[\epsilon]$ and $\max_i \rho_i \le R$, the risk of the detector $\widehat{\phi}_P$ in the Poisson case of problem $(\mathcal{D}_\rho)$ is $\le \epsilon$. When $\rho = \rho^P[\epsilon]$ and $\epsilon < 1/4$, the test associated with $\widehat{\phi}_P$ is κ_n-rate optimal with $\kappa_n = \kappa_n(\epsilon) := \frac{\ln(n/\epsilon^2)}{\ln(1/(4\epsilon))}$. Note that $\kappa_n(\epsilon) \to 2$ as $\epsilon \to +0$.*

4.2.3 Gaussian case

Now let the distribution P_μ of ω be normal with the mean μ and known variance $\sigma^2 > 0$, i.e. $\omega \sim \mathcal{N}(\mu, \sigma^2 I)$. For the sake of simplicity, assume also that the (convex and compact) nuisance set $\mathcal{V}$ is symmetric w.r.t. the origin. In such a case, the null hypothesis is

$$H_0 : \ \mu \in X := \{\mu = Av, \ v \in \mathcal{V}\}, \tag{37}$$

while the alternative $H_1(\rho)$ can be represented as the union, over $i = 1, ..., n$ and $\chi \in \{-1, 1\}$, of $2n$ hypotheses

$$H^{\chi,i}(\rho_i) : \quad \mu \in \chi Y_i(\rho_i) = \chi \{rAe[i] + Av : v \in \mathcal{V}, \rho_i \le r \le R\} \tag{38}$$

(note that $\{x = re[i] + v : v \in \mathcal{V}, -R \le r \le -\rho_i\} = -\{x = re[i] + v : v \in \mathcal{V}, R \ge r \ge \rho_i\}$ due to $\mathcal{V} = -\mathcal{V}$). Let $\mathrm{ErfInv}(\cdot)$ be the inverse error function: $\mathrm{Erf}(\mathrm{ErfInv}(s)) = s$, $0 < s < 1$. For $1 \le i \le n$ and $\chi \in \{-1, 1\}$, let us set (cf. section 2.3.1)

$$\rho_i^G(\epsilon) = \max_{\rho,r,u,v} \left\{ \rho : \begin{array}{l} \|A(u - re[i] - v)\|_2 \le \sigma \left[\mathrm{ErfInv}\left(\frac{\epsilon}{4n}\right) + \mathrm{ErfInv}\left(\frac{\epsilon}{2}\right)\right] \\ \chi r \ge \rho, \ u, v \in \mathcal{V} \end{array} \right\} \tag{$G_\epsilon^{i,\chi}$}$$

(the left hand side quantity clearly is independent of χ due to $\mathcal{V} = -\mathcal{V}$), and let

$$\begin{array}{rcl} \phi_{i,\chi}(\omega) & = & [A(u^{i,\chi} - r^{i,\chi}e[i] - v^{i,\chi})]^T\omega - \alpha_i, \\ \alpha_i & = & \lambda\,[A(u^{i,\chi} - r^{i,\chi}e[i] - v^{i,\chi})]^T[A(u^{i,\chi} + r^{i,\chi}e[i] + v^{i,\chi})], \\ \lambda & = & \frac{\mathrm{ErfInv}(\frac{\epsilon}{2})}{\mathrm{ErfInv}(\frac{\epsilon}{4n}) + \mathrm{ErfInv}(\frac{\epsilon}{2})}, \end{array} \tag{39}$$

where $u^{i,\chi}, v^{i,\chi}, r^{i,\chi}$ are the u, v, r-components of an optimal solution to $(G_\epsilon^{i,\chi})$ (of course, in fact $r^{i,1} = -r^{i,-1} = \rho_i^G(\epsilon)$, and, besides, we can assume w.l.o.g. that $u^{i,-1} = -u^{i,1}$, $v^{i,-1} = -v^{i,1}$). Finally, let

$$\rho^G[\epsilon] = [\rho_1^G(\epsilon); ...; \rho_n^G(\epsilon)], \quad \widehat{\phi}_G(\omega) = \min_{1 \le i \le n, \chi = \pm 1} \phi_{i,\chi}(\omega). \tag{40}$$

Properties of the test associated with detector $\widehat{\phi}_G$ can be described as follows:

Proposition 4.2 *Whenever $\rho \ge \rho^G[\epsilon]$ and $\max_i \rho_i \le R$, the risk of the test $\widehat{\phi}_G$ in the Gaussian case of problem $(\mathcal{D}_\rho)$ is $\le \epsilon$. When $\rho = \rho^G[\epsilon]$, the test is κ_n-rate optimal with*

$$\kappa_n = \kappa_n(\epsilon) := \frac{\mathrm{ErfInv}(\frac{\epsilon}{4n})}{2\mathrm{ErfInv}(\frac{\epsilon}{2})} + \tfrac{1}{2}.$$

Note that $\kappa_n(\epsilon) \to 1$ as $\epsilon \to +0$.

Remarks. The results of Propositions 4.1, 4.2 imply that testing procedures $\widehat{\phi}_G$ and $\widehat{\phi}_P$ are κ_n–rate optimal in the sense of the above definition with $\kappa_n \asymp \sqrt{\ln n}$ in the Gaussian case and $\kappa_n \asymp \ln n$ in the Poisson case. In particular, this implies that the detection rates of these tests are within a $\sqrt{\ln n}$ (resp., $\ln n$)–factor of the rate profile ρ^* of the "oracle detector" – (the best) detection procedure which "knows" the node $\gamma \in \Gamma$ at which an event may occur. This property of the proposed tests allows also for the following interpretation: consider the Gaussian

problem setting in which the standard deviation σ of noise is inflated by the factor κ_n. Then for every $i \in \{1, ..., 2n\}$ there is no test of hypothesis H_0 vs. $H^i(\rho_i)$ with risk $\leq \epsilon$, provided that $\rho_i < \rho_i^G(\epsilon)$.

Note that it can be proved that the price – the $\sqrt{\ln n}$–factor – for testing multiple hypotheses cannot be eliminated at least in some specific settings [22].

An important property of the proposed procedures is that they can be efficiently implemented – when the nuisance set $\mathcal{V}$ is computationally tractable (e.g., is a polyhedral convex set, an ellipsoid, etc.), the optimization problems $(G_\epsilon^{i,\chi})$, (P_ϵ^i) are well structured and convex and thus can be efficiently solved using modern optimization tools even in relatively large dimensions.

4.2.4 Numerical illustration: signal detection in the convolution model

We consider here the "convolution model" with observation $\omega = A(s + v) + \xi$, where $s, v \in \mathbb{R}^n$, and $\xi \sim \mathcal{N}(0, \sigma^2 I_m)$ with known $\sigma > 0$, and A is as follows. Imagine that we observe at m consecutive moments the output of a discrete time linear dynamical system with a given impulse response ("kernel") $\{g_k\}$ supported on a finite time horizon $k = 1, ..., T$. In this case, our observation $y \in \mathbb{R}^m$ is the linear image of n-dimensional "signal" x which is system's input on the observation horizon, augmented by the input at $T-1$ time instants preceding this horizon (that is, $n = m + T - 1$). A is exactly the $m \times n$ matrix (readily given by m and the kernel) of the just described linear mapping $x \mapsto y$.

We want to detect the presence of the signal $s = re[i]$, where $e[i]$, $i = 1, ..., n$, are some given vectors in $\mathbb{R}^n$. In other words, we are to decide between the hypotheses $H_0 : \mu \in A\mathcal{V}$ and $H_1(\rho) = \cup_{1 \leq i \leq n, \chi = \pm 1} H^{\chi,i}(\rho_i)$, with the hypotheses $H^{\chi,i}(\rho_i)$ defined in (38). The setup for our experiment is as follow: we use $g_k = (k+1)^2(T-k)/T^3$, $k = 0, ..., T-1$, with $T = 60$, and $m = 100$, which results in $n = 159$. The signatures $e[i]$, $1 \leq i \leq n$ are the standard basic orths in $\mathbb{R}^n$ or unit step functions: $e_k[i] = 1_{\{k \leq i\}}$, $k = 1, ..., n$, and the nuisance set $\mathcal{V}$ is defined as $\mathcal{V}_L = \{u \in \mathbb{R}^n : , |u_i - 2u_{i-1} - u_{i-2}| \leq L,\ i = 3, ..., n\}$, where L is experiment's parameter.

The goal of the experiment was to illustrate how large in the outlined problem is the (theoretically, logarithmic in n) "nonoptimality factor" $\kappa_n(\epsilon)$ of the detector $\widehat{\phi}_G$, specifically, how it scales with the risk ϵ. To this end, we have computed, for different values of ϵ, first, the "baseline profile" — the vector with the entries

$$\rho_i^*(\epsilon) = \max_{\rho, r, u, v} \{\rho : \|A(u - re[i] - v)\|_2 \leq 2\sigma\, \mathrm{ErfInv}(\epsilon/2), r \geq \rho,\ u, v \in \mathcal{V}\} \tag{41}$$

(cf. $(G_\epsilon^{i,1})$); $\rho_i^*(\epsilon)$ is just the smallest ρ for which the hypotheses H_0 and $H^{1,i}(\rho)$ can be distinguished with error probabilities $\leq \epsilon$ (recall that we are in the Gaussian case). Second, we computed the profile $\rho^G[\epsilon]$ of the test with detector $\widehat{\phi}_G$ underlying Proposition 4.2. The results are presented on figure 2. Note that for $\epsilon \leq 0.01$ we have $\rho^G(\epsilon)/\rho^*(\epsilon) \leq 1.3$ in the reported experiments.

Quantifying conservatism. While the baseline profile ρ^* establishes an obvious lower bound for the ρ-profile of any test in our detection problem, better lower bounds can be computed by simulations. Indeed, let

$$x_0^{i,\chi} = \chi u^i,\ x_1^{i,\chi} = \chi \rho_i e[i] + v^i,\ i = 1, ..., n,\ \chi \in \{-1, 1\},$$

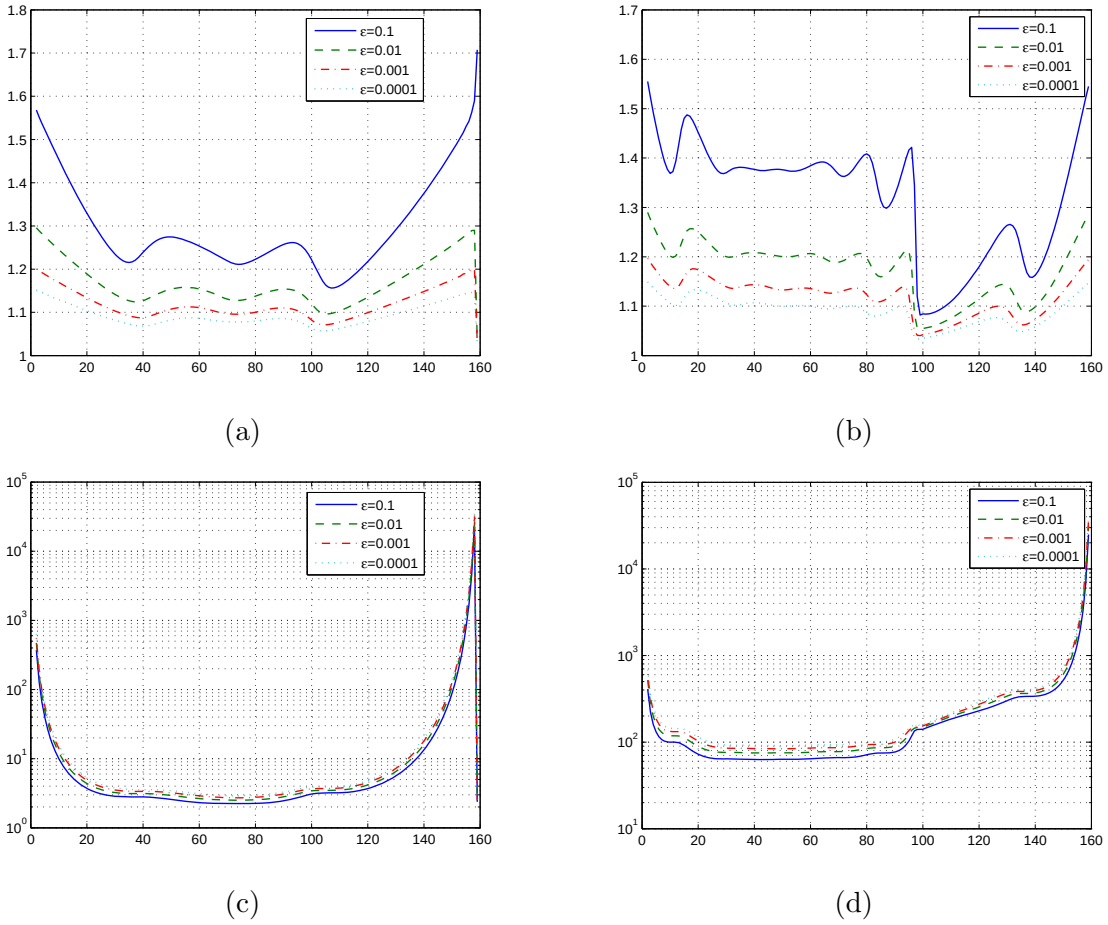

Figure 2: The left pane (plots (a) and (c)) represents the experiment with "step" signals, the right pane (plots (c) and (d)) corresponds to the experiment with the signals which are proportional to basis orths. Nuisance parameter is set to $L = 0.1$ and $\sigma = 1$ in both experiments. Plots (a) and (b): the value of $\rho^G[\epsilon]/\rho^*[\epsilon]$ for different values of ϵ; plots (c) and (d): corresponding rate profiles (logarithmic scale).

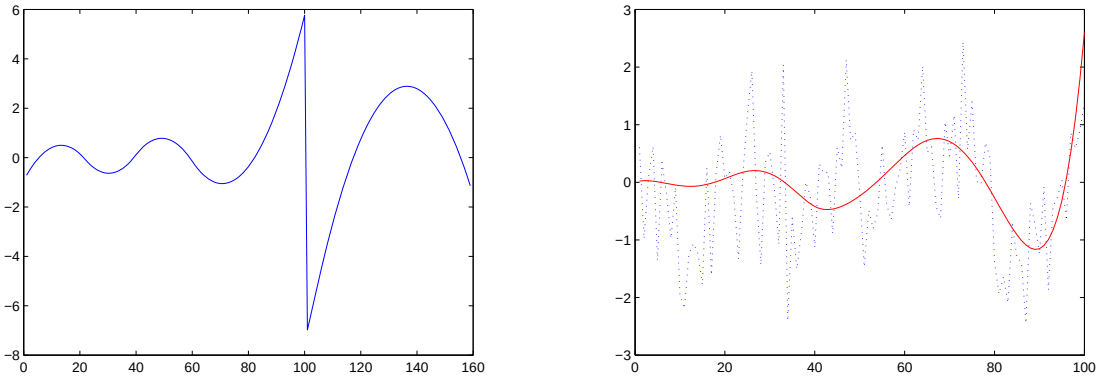

Figure 3: "Hard to detect" signal $\rho_i^G(\epsilon)e[i] + v^{i,1} - u^{i,1}$, where $\rho_i^G(\epsilon)$, $v^{i,1}$ and $u^{i,1}$ are components of an optimal solution to $(G_\epsilon^{i,\chi})$ with $\epsilon = 0.05$ and $i = 100$ (left plot), and its image Ax with a noisy observation (right plot). Experiment with "step" useful signals, nuisance parameter $L = 0.1$ and $\sigma = 1$.

where v^i and u^i are some vectors in $\mathcal{V}$. It is clear that the optimal risk in the problem of distinguishing H_0 and $H_1(\rho) = \bigcup_{i=1}^n H^{\chi,i}(\rho_i)$ (cf. (37) and (38)) is lower bounded by the risk of distinguishing

$$\bar{H}_0 : \ \mu \in \{Ax_0^{i,\chi}, i = 1, ..., n, \ \chi \in \{-1, 1\}\}, \text{ and } \bar{H}_1(\rho) : \ \mu \in \{Ax_1^{i,\chi}, i = 1, ..., n, \ \chi \in \{-1, 1\}\},$$

which, in its turn, is lower bounded by the risk of distinguishing of the hypothesis $\tilde{H}_0 : \ \mu = 0$ from the alternative

$$\tilde{H}_1(\rho) : \ \mu \in \{Az^{i,\chi}, \ z^{i,\chi} = x_1^{i,\chi} - x_0^{i,\chi} = \chi(\rho_i e[i] + v^i - u^i), \ i = 1, ..., n, \ \chi \in \{-1, 1\}\}.$$

On the other hand, the latter risk is clearly bounded from below by the risk of the Bayesian test problem as follows:

$$(\mathcal{D}_\rho^\nu) \quad \begin{array}{l} \text{Given } \rho = [\rho_1; ...; \rho_n] > 0, \text{ test the hypothesis } H_0 : \ \mu = 0 \text{ against the} \\ \text{alternative } H_1^\nu(\rho) : \ \mu = \chi A(\rho_i e[i] + v^i - u^i) \text{ with probability } \nu_{\chi i} \\ \text{where } v^i, u^i \in \mathcal{V}, \text{ and } \nu \text{ is a probability on } \{\chi i\}, \ i = 1, ..., n, \ \chi \in \{-1, 1\}. \end{array}$$

We conclude that the risk of deciding between H_0 and $H_1(\rho)$ may be lower bounded by the risk of the optimal (Bayesian) test in the Bayesian testing problem $(\mathcal{D}_\rho^\nu)$. Note that we are completely free to choose the distribution ν and the points $u^i, v^i \in \mathcal{V}$, $i = 1, ..., n$. One can choose, for instance, $v^{\cdot,\chi}$ and $v^{\cdot,\chi}$ as components of an optimal solution to (41) and a uniform on $\{\pm 1, ..., \pm n\}$ prior probability ν. Let us consider the situation where the matrix A is an $n \times n$ Toeplitz matrix of periodic convolution on $\{1, ..., n\}$ with kernel g, $g_k = (\frac{k}{T})^2(1 - \frac{k}{T})$, $k = 1, .., T$, signatures $e[i] = e_{\cdot - i}$ are the shifts of the same signal $e_k = k/n$, $k = 1, ..., n$, and the nuisance set

$$\mathcal{V}_L = \{u \in \mathbb{R}^n : , |u_i - 2u_{i-1 \bmod n} - u_{i-2 \bmod n}| \le L, \ i = 1, ..., n\}$$

is symmetric and shift-invariant. Let us fix $\epsilon > 0$ and choose $v^i = -u^i$ as components of an optimal solution to the corresponding optimization problem $(G_\epsilon^{i,\chi})$. Because of the shift-invariance of the problem setup the optimal values $\rho_i^*(\epsilon)$ and $\rho_i^G(\epsilon)$ do not depend on i and are

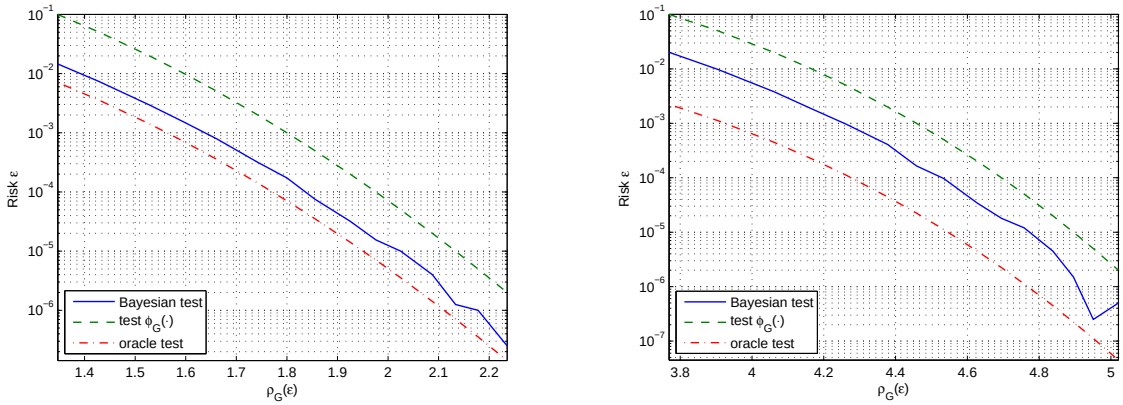

Figure 4: Estimated risk of the Bayes test as a function of test rate $\rho_G(\epsilon)$, compared to the risk of the baseline test and that of the simple test with data (39) ($L = 0.01$ and $\sigma = 1$). Simulation for $n = 100$ (left plot) and $n = 1000$ (right plot).

equal to the same $\rho^*(\epsilon)$ and, respectively, $\rho^G(\epsilon)$, and all v^i are the shifts of the same $v \in \mathbb{R}^n$. In this case the risk of the Bayesian test corresponding to the uniform on $\{\pm 1, ..., \pm n\}$ prior distribution ν is a lower bound of the optimal risk for the corresponding detection problem $(\mathcal{D}_\rho)$.

On figure 4 we present the results of two simulation for $n = 100$ and $n = 1000$, the value $L = 0.01$ of the parameter of the nuisance class, and $\sigma = 1$. For different values of ϵ we have first computed corresponding rates $\rho^*(\epsilon)$ and $\rho^G(\epsilon)$, as well as components $v^i = -u^i$ of the optimal solution (recall that due to the shift-invariance of the problem, $v^i_k = v^1_{k-i+1 \bmod n}$). Then an estimation of the risk of the Bayesian test with the uniform prior is computed over $N = 10^7$ random draws. Note that already for $\epsilon = 0.01$ rate $\rho^G(\epsilon)$ of the simple test is only 7% higher than the corresponding Bayesian lower bound for $n = 1000$ (15% for $n = 100$).

4.2.5 Numerical illustration: signal identification in the convolution model

The purpose of the experiment we report on in this section is to illustrate an application of the approach to multiple hypotheses testing presented in section 3.2.1. The experiment in question is a modification of that described in section 4.2.4, the setup is as follows. On time horizon $t = 1, ..., m$, we observe the output, contaminated by noise, of a discrete-time linear dynamic system with "memory" T (that is, the impulse response g is zero before time 0 and after time $T - 1$). The input x to the system is an impulse of amplitude $\geq \rho > 0$ (ρ is known) at unknown time τ known to satisfy $-T + 2 \leq \tau \leq m$. Setting $n = m + T - 1$, our observation is

$$\omega = [\omega_1; ...; \omega_m] = Ax + \xi, \ \xi \sim \mathcal{N}(0, I_m),$$

with $m \times n$ matrix A readily given by the impulse response g. We have n hypotheses about x, the i-th of them stating that $x \in X_i = \{x = re_i, r \geq \rho\}$, where e_i, $i = 1, ..., n$, are the standard basic orths in $\mathbb{R}^n$. Given an observation, we want to decide to which of the sets $X_1, ..., X_n$ the

actual input belongs, that is, we need to distinguish between n hypotheses $H_1, ..., H_n$ on the distribution of ω, with H_i stating that this distribution is $\mathcal{N}(Ax, I_m)$ for some $x \in X_i$.

The problem can be processed as follows. Let us choose two nonnegative integers μ ("margin") and ν ("resolution"), and imagine that we do not care much about distinguishing between the "boundary hypotheses" H_i (those with $i \leq \mu$ and with $i \geq n - \mu + 1$) and all other hypotheses, same as we do not care much about distinguishing between "close to each other" hypotheses H_i and H_j, those with $|i - j| \leq \nu$. What we do care about is not to miss the true hypothesis and to reject any non-boundary hypothesis which is not close to the true one. Note that when $\mu = \nu = 0$, we "care about everything;" this, however, could require large amplitude ρ in order to get a reliable test, since the impulses at times t close to the endpoints of the time segment $-T + 2 \leq t \leq m$ could be poorly observed, and impulses at close to each other time instants could be difficult to distinguish. Operating with positive margins and/or resolutions, we, roughly speaking, sacrifice the "level of details" in our conclusions in order to make these conclusions reliable for smaller values of the amplitude ρ.

With the approach developed in section 3.2.1, our informally described intentions can be formalized as follows. In the terminology and notation of section 3.2.1, let us define the set $\mathcal{C}$ of pairs (i, j), $1 \leq i, j \leq n$, $i \neq j$, i.e., the pairs with "H_j close to H_i," as follows:

- for a "boundary hypothesis" H_i (one with $1 \leq i \leq \mu$ or $n - \mu + 1 \leq i \leq n$), every other hypothesis H_j is close to H_i;
- for a "non-boundary hypothesis" H_i (one with $1 + \mu \leq i \leq n - \mu$), close to H_i hypotheses H_j are those with $1 \leq |i - j| \leq \nu$.

Detectors $\phi_{ij}(\omega)$ we intend to use are the Gaussian log-likelihood detectors

$$\begin{array}{c} \phi_{ij}(\omega) = \frac{1}{2}[\xi_{ij} - \eta_{ij}]^T\omega + \frac{1}{4}[\eta_{ij}^T\eta_{ij} - \xi_{ij}^T\xi_{ij}], \\ \xi_{ij} = Ax_{ij},\ \eta_{ij} = Ay_{ij}, [x_{ij} = y_{ij}] = \mathrm{argmin}_{r,s}\{\|rAe_i - sAe_j\|_2 : r \geq \rho, s \geq \rho\}, \end{array} \tag{42}$$

which allows to specify the quantities ϵ_{ij} in (27) as

$$\epsilon_{ij} = \exp\{-(\xi_{ij} - \eta_{ij})^T(\xi_{ij} - \eta_{ij})/8\}, \tag{43}$$

see section 2.3.1.

Applying the construction from section 3.2.1, we arrive at a risk bound ϵ and a test which, given an observation ω, accepts some of the hypotheses H_i, ensuring the following. Let the true hypothesis be H_{i_*}. Then (all probabilities are taken according to the distribution specified by H_{i_*})

A. The probability for H_{i_*} to be rejected by the test is at most ϵ;

B. The probability of the event that the list of accepted hypotheses contains a hypothesis H_j such that both H_j is *not* close to H_{i_*} and H_{i_*} is not close to H_j is at most ϵ.

   Note that with our definition of closeness, the latter claim implies that when H_{i_*} is not a boundary hypotheses, the probability for the list of accepted hypotheses to contain a *non-boundary* hypothesis H_j with $|i - j| > \nu$ is at most ϵ.

The outlined model demonstrates the potential of *asymmetric* closeness: when a boundary hypothesis is difficult to distinguish from other hypotheses, it is natural to declare all these

	$\nu=0$	$\nu=1$	$\nu=2$	$\nu=3$
$\mu=0$	276.0(+40.1%) 1.00	71.0(+40.0%) 1.00	31.5(+40.4%) 1.00	18.1(+44.4%) 1.03
$\mu=1$	133.2(+40.5%) 1.88	48.0(+40.5%) 1.48	23.6(+40.3%) 1.33	14.1(+40.5%) 1.25
$\mu=2$	102.0(+40.2%) 1.44	36.8(+40.0%) 1.93	19.4(+40.3%) 1.64	11.9(+40.1%) 1.48
$\mu=3$	77.5(+40.1%) 1.33	29.8(+40.0%) 1.61	16.3(+40.3%) 1.94	10.4(+40.1%) 1.70

Table 1: Identifying signals in the convolution model. In a cell, top: $\rho(\mu,\nu)$ and excess $\rho(\mu,\nu)/\underline{\rho}(\mu,\nu)-1$ (in brackets, percents); bottom: $\widetilde{\rho}(\mu,\nu)/\rho(\mu,\nu)$.

hypotheses to be close to the boundary one. On the other hand, there are no reasons to declare a boundary hypothesis to be close to a well identifiable "inner" hypothesis.

As we have seen in section 3.2.1, given ρ, the risk ϵ can be efficiently computed via convex optimization, and we can use this efficient computation to find the smallest amplitude ρ for which ϵ takes a given target value ε. This is what was done in the numerical experiment we are about to report. In this experiment, we used $T=m=16$ (i.e., the number of hypotheses n was 31), and the impulse response was similar to the one reported earlier in this section, namely the nonzero entries in g were

$$g_t=\alpha(t+1)^2(T-t),\ 0\le t\le T-1,$$

while α was selected to ensure $\max_t g_t=1$. For various values of margins μ and resolutions ν, we computed the minimal amplitude $\rho=\rho(\mu,\nu)$ which still allowed for our test to guarantee risk $\epsilon\le 0.01$. The results are presented in table 1. A simple lower bound $\underline{\rho}(\mu,\nu)$ on the smallest ρ such that there exists "in the nature" a test capable to ensure A and B with $\epsilon=0.01$, amplitudes of impulses being ρ, may be constructed by lower bounding the probability of a union of events by the largest among the probabilities of these events. In the table we present, along with the values of $\rho(\cdot,\cdot)$, the "excess value" $\rho(\mu,\nu)/\underline{\rho}(\mu,\nu)-1$. Observe that while $\rho(\mu,\nu)$ itself strongly depends on the margin μ, the excess is nearly independent of μ and ν. Of course, 40% excess is unpleasantly large; note, however, that the lower bound $\underline{\rho}$ definitely is optimistic. In addition, this "overly pessimistic" excess decreases as the target value of ϵ decreases; what was 40% for $\varepsilon=0.01$, becomes 26% for $\varepsilon=0.001$ and 19% for $\varepsilon=1.e\text{-}4$.

In the reported experiment, along with identifying $\rho(\cdot,\cdot)$, we were interested also in the effect of optimal shifts $\phi_{ij}(\cdot)\mapsto\phi_{ij}(\cdot)-\bar{\alpha}_{ij}$, see section 3.2.1. To this end we compute the smallest $\rho=\widetilde{\rho}(\mu,\nu)$ such that the version of our test utilizing $\alpha_{ij}\equiv 0$ is capable to attain the risk $\varepsilon=0.01$. Table 1 presents, along with other data, the ratios $\widetilde{\rho}(\mu,\nu)/\rho(\mu,\nu)$ which could be considered as quantifying the effect of shifting the tests. We see that the effect of the shift is significant when the margin μ is positive.

4.3 Testing from indirect observations

4.3.1 Problem description

Let $\mathcal{F}$ be a class of cumulative distributions on $\mathbb{R}$. Suppose that for $\ell = 1, ..., L$, we are given K_ℓ independent realizations of random variable ζ^ℓ. We assume that the c.d.f. F_{ζ^ℓ} of ζ^ℓ is a linear transformation of unknown c.d.f. F_ξ of "latent" random variable ξ, $F_\xi \in \mathcal{F}$. In this section we consider two cases of the sort; in both of them, η^ℓ is an independent of ξ random variable ("nuisance") with known c.d.f. F_{η^ℓ}. In the first case ("deconvolution model"), $\zeta^\ell = \xi + \eta^\ell$, so that the distribution of ζ^ℓ is $F_{\zeta^\ell}(t) = \int_{\mathbb{R}} F_\xi(t-s)dF_{\eta^\ell}(s)$. In the second case ("trimmed observations"), observations are trimmed: $\zeta^\ell = \max\{\xi, \eta^\ell\}$, so that $F_{\zeta^\ell}(t) = F_\xi(t)F_{\eta^\ell}(t)$.

We consider here the testing problem where our objective is to test, for given $t \in \mathbb{R}$, $\alpha \in (0,1)$ and $\rho > 0$, the hypotheses[12]

$$H_1 : F_\xi(t) < \alpha - \rho \text{ and } H_2 : F_\xi(t) > \alpha + \rho \qquad (C_{\alpha,t}[\rho])$$

given observations ζ_k^ℓ, $k = 1, ..., K_\ell$, $\ell = 1, ..., L$.

Under minor regularity conditions on F_{η^ℓ} and F_ξ, $(C_{\alpha,t}[\rho])$ may be approximated by the discrete decision problem as follows. Let ξ be a discrete random variable with unknown distribution x known to belong to a given closed convex subset $\mathcal{X}$ of n-dimensional probabilistic simplex. We want to infer about x given indirect observations of ξ obtained by L different "observers": the observations ω_i^ℓ, $i = 1, ..., K_\ell$ of ℓ-th observer are independent realizations of random variable ω^ℓ taking values $1, ..., m_\ell$ with distribution $\mu^\ell = A^\ell x$, where A^ℓ is a known stochastic matrix. For instance, when ξ takes values $1, ..., n$ and $\omega^\ell = \xi + \eta^\ell$ with nuisance η^ℓ taking values $1, ..., n_\ell$ and distribution u^ℓ, A^ℓ is $(n_\ell + n - 1) \times n$ matrix, and the nonzero entries of the matrix are given by $A^\ell_{ij} = u^\ell_{i-j+1}$, $1 \le j \le i \le j + n_\ell - n$. We assume in the sequel that $A^\ell x > 0$ whenever $x \in \mathcal{X}$, $1 \le \ell \le L$.

Let $g(x) = g^T x$, $g \in \mathbb{R}^n$, be a given linear functional of the distribution x. Given α and $\rho > 0$, our goal is to decide on the hypotheses about the distribution x of ξ

$$H_1[\rho] : x \in \mathcal{X},\ g(x) \le \alpha - \rho, \quad H_2[\rho] : x \in \mathcal{X},\ g(x) \ge \alpha + \rho. \qquad (\mathcal{D}_{g,\alpha}[\rho])$$

given observations $\omega^1, ..., \omega^\ell$. We denote by $\rho_{\max}$ the largest ρ for which both these hypotheses are nonempty, and assume from now on that $\rho_{\max} > 0$ (as far as our goal is concerned, this is the only nontrivial case). Now let us fix $0 < \epsilon < 1$ and, given a decision rule $T(\cdot)$, let us denote $\rho_T[\epsilon]$ the smallest $\rho \ge 0$ such that the risk of the rule $T(\cdot)$ in the problem $(\mathcal{D}_{g,\alpha}[\rho])$ does not exceed ϵ. We refer to $\rho_T[\epsilon]$ as the ϵ-resolution of $T(\cdot)$ and denote by $\rho^*[\epsilon] = \inf_{T(\cdot)} \rho_T[\epsilon]$ ("ϵ-rate") the best ϵ-resolution achievable in our problem. Our goal is given ϵ, to design a test with ϵ-resolution close to $\rho^*[\epsilon]$.

The resulting observation scheme fits the definition of the direct product of Discrete observation schemes of section 2.4.2 – we have $K = \sum_{\ell=1}^L K_\ell$ "simple" (or L K_ℓ-repeated) Discrete observation schemes, the k-th scheme yielding the observation ω_k, $k = 1, ..., K$, of one of L types.

Given an $\epsilon \in (0,1)$, we put

$$\rho[\epsilon] = \max_{x,y,r} \left\{ r : \begin{array}{l} \sum_{\ell=1}^L K_\ell \ln\left(\sum_{i=1}^{m_\ell} \sqrt{[A^\ell x]_i [A^\ell y]_i}\right) \ge \ln \epsilon, \\ x, y \in \mathcal{X},\ g(x) \le \alpha - r,\ g(y) \ge \alpha + r. \end{array} \right\} \qquad (44)$$

[12] A related problem of *estimation* of the c.d.f. F_ξ in the deconvolution model, a special case of linear functional estimation [18, 19, 32], have received much attention in the statistical literature (see, e.g., [21, 47, 20, 16] and [38, Section 2.7.2] for a recent review of corresponding contributions).

Clearly, $0 \le \rho[\epsilon] \le \rho_{\max}$ due to $\rho_{\max} > 0$. *We assume from now on that* $\rho[\epsilon] < \rho_{\max}$. Let now $\rho \in [\rho[\epsilon], \rho_{\max}]$. Consider the optimization problem

$$\mathrm{Opt}[\rho] = \max_{x,y} \left\{ \Psi(x,y) : \begin{array}{l} \Psi(x,y) = \sum_{\ell=1}^{L} K_\ell \ln\left(\sum_{i=1}^{m_\ell} \sqrt{[A^\ell x]_i [A^\ell y]_i}\right), \\ x, y \in \mathcal{X},\ g(x) \le \alpha - \rho,\ g(y) \ge \alpha + \rho. \end{array} \right\}. \qquad (F_{g,\alpha}[\rho])$$

This problem is feasible (since $\rho \le \rho_{\max}$) and thus solvable, and from $\rho \ge \rho[\epsilon]$ and $\rho[\epsilon] < \rho_{\max}$ it easily follows (see item 1^0 in the proof of Proposition 4.3) that $\mathrm{Opt}[\rho] \le \epsilon$. Let (x_ρ, y_ρ) be an optimal solution. Consider a simple test $\widehat{T}_\rho$ given by the detector $\widehat{\phi}(\cdot)$,

$$\widehat{\phi}(\omega) = \widehat{\phi}_\rho(\omega) := \sum_{k=1}^{K} \phi_k(\omega_k),\ \ \phi_k(\omega_k) = \tfrac{1}{2} \ln\left([A^{\ell(k)} x_\rho]_{\omega_k} / [A^{\ell(k)} y_\rho]_{\omega_k} \right), \qquad (45)$$

with $\ell(k)$ uniquely defined by the relations

$$\sum_{\ell < \ell(k)} K_\ell < k \le \sum_{\ell \le \ell(k)} K_\ell.$$

We have the following simple corollary of Proposition 2.2:

Proposition 4.3 *Assume that* $\rho_{\max} > 0$ *and* $\rho[\epsilon] < \rho_{\max}$, *and let* $\epsilon \in (0, 1/4)$. *Then*

$$\rho[\epsilon] \le \vartheta(\epsilon)\rho^*[\epsilon],\ \ \vartheta(\epsilon) = \frac{2\ln(1/\epsilon)}{\ln[1/(4\epsilon)]}. \qquad (46)$$

In other words, there is no decision rule in the problem $(\mathcal{D}_{g,\alpha}[\rho])$ *with the risk* $\le \epsilon$ *if* $\rho < \rho[\epsilon]/\vartheta(\epsilon)$.

On the other hand, when $\rho \in [\rho[\epsilon], \rho_{\max}]$, *the risk of the simple test* $\widehat{\phi}_\rho$ *in the problem* $(\mathcal{D}_{g,\alpha}[\rho])$ *is* $\le \exp(\mathrm{Opt}[\rho]) \le \epsilon$.

Note that $\vartheta(\epsilon) \to 2$ as $\epsilon \to 0$. Under the premise of Proposition 4.3, the test associated with detector $\widehat{\phi}_{\rho[\epsilon]}(\cdot)$ is well defined and distinguishes between the hypotheses $H_1[\rho[\epsilon]]$, $H_2[\rho[\epsilon]]$ with risk $\le \epsilon$. We refer to the quantity $\rho[\epsilon]$ as to *resolution* of this test.

4.3.2 Numerical illustration

We present here some results on numerical experimentation with the testing problem $(C_{\alpha,t}[\rho])$. For the sake of simplicity, we suppose that the distributions with c.d.f.'s from $\mathcal{F}$ are supported on $[-1, 1]$. We start with an appropriate discretization of the continuous problem.

Discretizing continuous model.

1. Let $n \in \mathbb{Z}_+$, and let $-1 = a_0 < a_1 < a_2 < ... < a_n = 1$ be a partition of $(-1, 1]$ into n intervals $I_i = (a_{i-1}, a_i]$, $i = 1, ..., n$. We associate with a c.d.f. $F \in F$ the n-dimensional probabilistic vector $x = x[F]$ with the entries $x_k = \mathrm{Prob}_{\xi \sim F}\{\xi \in I_k\}$ and $\bar{a}_k = (a_{k-1} + a_k)/2$, the central point of I_k, $k = 1, ..., n$, and denote by $\mathcal{F}_n$ the image of $\mathcal{F}$ under the mapping $F \mapsto x[F]$.

2. We build somehow a convex compact subset $\mathcal{X} \supset \mathcal{F}_n$ of the n-dimensional probabilistic simplex.

3. Depending on the observation scenario, we act as follows.

   (a) *Deconvolution problem:* ζ^ℓ satisfy $\zeta^\ell = \xi + \eta^\ell$. Let $0 < \delta < 1$ (e.g., $\delta = K_\ell^{-1}$), $m_\ell \in \mathbb{Z}_+$, and let

   $$b_1^\ell = a_0 + q_{\eta^\ell}(\delta),\ b_{m_\ell - 1}^\ell = a_n + q_{\eta^\ell}(1-\delta),$$

   where $q_{\eta^\ell}(p)$ is the p-quantile of η_ℓ. Note that $\mathrm{Prob}\{\zeta^\ell \notin [b_1^\ell, b_{m_\ell-1}^\ell]\} \le 2\delta$. Let now $-\infty = b_0^\ell < b_1^\ell < b_2^\ell < ... < b_{m_\ell-1}^\ell < b_m^\ell = \infty$ be a partition of $\mathbb{R}$ into m_ℓ intervals $J_i^\ell = (b_{i-1}^\ell, b_i^\ell]$, $i = 1, ..., m_\ell - 1$, $J_{m_\ell} = (b_{m_\ell-1}^\ell, \infty)$. We put $\mu_i^\ell = \mathrm{Prob}\{\zeta \in J_i\}$, $i = 1, ..., m_\ell$ and define the $m_\ell \times n$ matrix stochastic matrix $A^\ell = (A_{jk}^\ell)$ with elements

   $$A_{ij}^\ell = \mathrm{Prob}\left\{\bar{a}_j + \eta^\ell \in J_i\right\},$$

   the approximations of conditional probabilities $\mathrm{Prob}\{\zeta^\ell \in J_i | \xi \in I_j\}$.

   (b) *Trimmed observations:* $\zeta^\ell = \max\{\xi, \eta^\ell\}$. We partition $\mathbb{R}$ into $m_\ell = n+1$ intervals, I_i, $i = 1, ..., n$ as above and an "infinite bin" $I_{n+1} = (a_n, a_{n+1} = \infty)$. We put $\mu_i^\ell = \mathrm{Prob}\{\zeta \in J_i\}$, $i = 1, ..., m_\ell$ and define the $m_\ell \times n$ matrix A^ℓ with elements

   $$A_{ij}^\ell = \delta_{ij}\mathrm{Prob}\{\eta^\ell \le a_j\} + 1_{\{i>j\}}\mathrm{Prob}\{\eta^\ell \in I_i\},$$

   where $\delta_{ij} = 1$ if $i = j$ and zero otherwise, which are the estimates of the probability of ζ^ℓ to belong to I_i, given that $\xi \in I_j$.

4. We denote $g = g(t) \in \mathbb{R}^n$, with entries $g_i = 1_{\{\bar{a}_i \le t\}}$, $i = 1, \ldots, n$, so that $g^T x$ is an approximation of $F(t)$.

5. Finally, we consider discrete observations $\omega_k^\ell \in \{1, ..., m_\ell\}$,

   $$\omega_k^\ell = i\, 1_{\{\zeta_k^\ell \in J_i^\ell\}}\ \ k = 1, ..., K^\ell,\ \ell = 1, ..., L.$$

We have specified the data of a testing problem of the form $(\mathcal{D}_{g,\alpha}[\rho])$. Note that the discrete observations we end up with are deterministic functions of the "true" observations ζ^ℓ, so that a test for the latter problem induces a test for the problem of interest $(C_{\alpha,t}[\rho])$. When distributions from $\mathcal{F}$, same as distributions of the nuisances η^ℓ, possess some regularity, and the partitions (I_i) and (J_i) are "fine enough", the problem $(\mathcal{D}_{g,\alpha}[\rho])$ can be considered as a good proxy of the problem of actual interest.

Simulation study. We present results for three distributions of the nuisance:

(i) Laplace distribution $\mathcal{L}(\mu, a)$ (i.e., the density $(2a)^{-1}e^{-|x-\mu|/a}$) with parameter $a = \frac{1}{2}$ and $\mu = 0$;

(ii) distribution $\Gamma(0, 2, 1/(2\sqrt{2}))$ with the location 0, shape parameter 2 and the scale $\frac{1}{2\sqrt{2}}$ (the standard deviation of the error is equal to 0.5).[13]

[13] Recall that Γ-distribution with parameters μ, α, θ has the density $[\Gamma(\alpha)\theta^\alpha]^{-1}(x-\mu)^{\alpha-1}\exp\{-(x-\mu)/\theta\}1_{\{x\ge\mu\}}$.

(iii) mixture of Laplace distributions $\frac{1}{2}\mathcal{L}(-1,\frac{1}{2})+\frac{1}{2}\mathcal{L}(1,\frac{1}{2})$.

The interval $[-1,1]$ was split into $n=100$ bins of equal lengths. The discretized distributions $x=x[F]$, $F\in\mathcal{F}$, are assumed to have bounded second differences, specifically, when denoting h the length of the bin,

$$|x_{i+1}-2x_i+x_{i-1}|\le h^2\mathcal{L},\ i=2,...,n-1;$$

in the presented experiments, $\mathcal{X}$ is comprised of all probabilistic vectors satisfying the latter relation with $\mathcal{L}=0.4$.

On figures 5 and 6 we present details of the test in the deconvolution model with $L=2$ observers. Each observer acquires K_ℓ noisy observations ζ_k^ℓ, $k=1,...,K_\ell$. The distribution of the nuisance is mixed Laplace for the first observer and $\Gamma(0,2,1/2/\sqrt{(2)})$ for the second observer. The discretized model has the following parameters: the observation spaces $\Omega_\ell=\mathbb{R}$, $\ell=1,2$ of each of 2 K_ℓ–repeated observation schemes were split into $m_\ell=102$ "bins": we put $b_1^\ell=-1+q_{\eta^\ell}([K^\ell]^{-1})$ and $b_{100}^\ell=1+q_{\eta^\ell}(1-[K^\ell]^{-1})$, and split the interval $(b_1^\ell,b_{100}^\ell]$ into 100 equal length bins; then we add two bins $(-\infty,b_1^\ell]$ and (b_{100}^ℓ,∞).

On figure 7 we present simulation results for the experiments with trimmed observations. Here $L=1$, the observations are $\omega_k=\max[\xi_k,\eta_k]$, $1\le k\le K$, with the $\mathcal{L}(0,\frac{1}{2})$ nuisances η_k. The partition of the support $[-1,1]$ of ξ is the same as in the deconvolution experiments, and the observation domain was split into $m=101$ bins – 100 equal length bins over the segment $[-1,1]$ and the bin $(1,\infty)$.

Quantifying conservatism. When building the test $\widehat{T}_\rho$ deciding on the hypotheses $H_\imath[\rho]$, $\imath=1,2$ (see $(\mathcal{D}_{g,\alpha}[\rho])$) via K observations $\omega^K=(\omega_1,...,\omega_K)$, we get, as a byproduct, two probability distributions $x_\rho\in\mathcal{F}$, $y_\rho\in\mathcal{F}$, of the latent random variable ξ, see (45). These distributions give rise to two simple hypotheses, $\overline{H}_1$, $\overline{H}_2$, on the distribution of observation ω^K, stating that these observations come from the distribution x_ρ, resp., y_ρ, of the latent variable. The risk of any test deciding on the two simple hypotheses $\overline{H}_1$, $\overline{H}_2$, the observation being ω^K, is lower-bounded by the quantity $\widehat{\epsilon}[K]=\sum_{\omega^K}\min[p_1^K(\omega^K),p_2^K(\omega^K)]$, where $p_i^K(\omega^K)$ is the probability to get an observation ω^K under hypothesis $\overline{H}_i$, $i=1,2$. The quantity $\widehat{\epsilon}[K]$, which can be estimated by Monte-Carlo simulation, by its origin is a lower bound on the risk of a whatever test deciding, via ω^K, on the composite "hypotheses of interest" $H_\imath[\rho]$, $\imath=1,2$. We can compare this lower risk bound with the upper bound $\epsilon[K]=\exp\{\mathrm{Opt}[\rho]\}$ on the risk of the test $\widehat{T}_\rho$, see $(F_{g,\alpha}[\rho])$, and thus quantify the conservatism of the latter test. The setup of the related experiments was completely similar to the one in the just reported experiments, with the Laplace distribution $\mathcal{L}(0,1/2)$ of the nuisance and with $n=500$ and $m=1002$ bins in the supports of ξ and of ω, respectively. We used $t=0$, $\alpha=0.5$, and 2×10^6 Monte-Carlo simulations to estimate $\widehat{\epsilon}[K]$. In our experiments, given a number of observations K and a prescribed risk level $\epsilon\in\{0.1,0.01,0.001,0.0001\}$, the parameter ρ of the test $\widehat{T}_\rho$ was adjusted to ensure $\epsilon[K]=\epsilon$; specifically, we set $\rho=\rho[\epsilon]$, see (44). The results are presented in table 2.

Recall that by Proposition 2.2 we have $\epsilon[K']\le(\epsilon[K])^{K'/K}$ when $K'\ge K$, so that the ratios $r[k]=\ln(\widehat{\epsilon}[K])/\ln(\epsilon[K])$ presented in the table upper-bound the nonoptimality of $\widehat{T}_\rho$ in terms of the number of observations required to achieve the risk $\widehat{\epsilon}[K]$: for the "ideal" test, at least K observations are required to attain this risk, and for the test $\widehat{T}_\rho$ – at most $\lfloor r[k]K\rfloor$ observations are enough. The data in table 2 show that the ratios $r[K]$ in our experiments newer exceeds 1.82 and steadily decrease when $\epsilon[K]$ decreases.

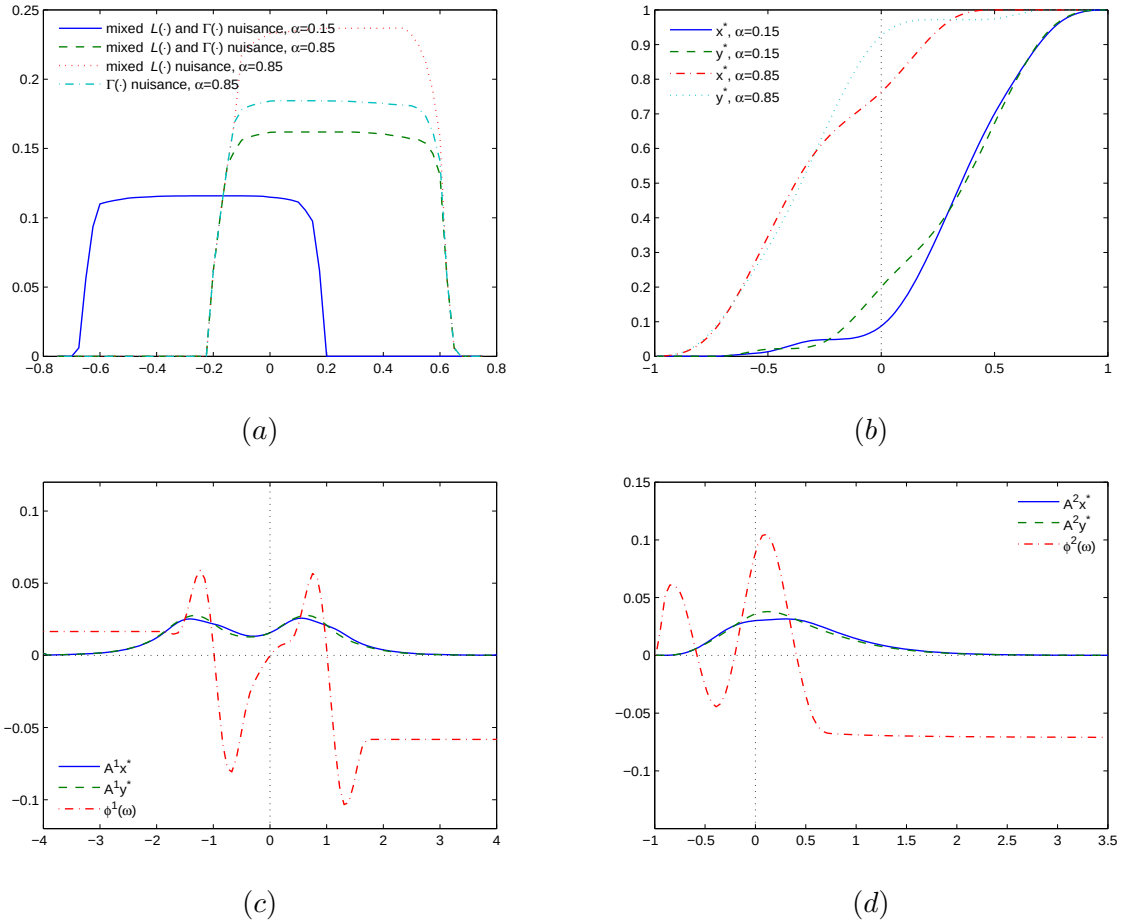

Figure 5: Deconvolution experiment, $K_\ell = 1000$, $k = 1, 2$, $\epsilon = 0.05$. In the upper row: (a) resolution of the simple test as a function of $t \in [-1, 1]$; (b) c.d.f. of the "difficult to test" distributions x^* and y^*, corresponding optimal solutions to $(F_{g,\alpha}[\rho])$ for $g = g(0)$ (testing hypotheses about $F(0)$). Bottom row: convolution images of optimal solutions to $(F_{g,\alpha}[\rho])$, $\alpha = .85$ and $g = g(0)$, and corresponding detector ϕ: (c) convolution with mixed Laplace distribution, (d) convolution with $\Gamma(\cdot)$ distribution.

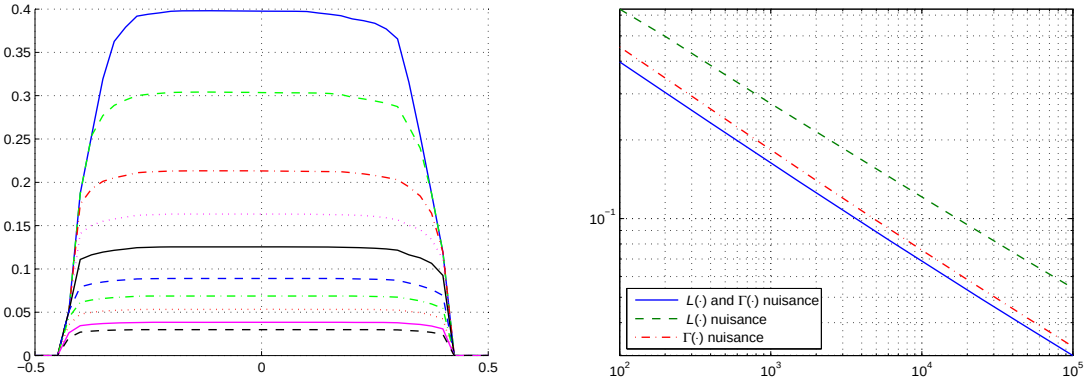

Figure 6: Deconvolution experiment, $\epsilon = 0.05$, $\alpha = 0.5$; $K_\ell = [100, 200, 500, 1000, ..., 100\,000]$, $\ell = 1, 2$. On the left: resolution of the simple test as a function of $t \in [-1, 1]$ for different K^ℓ, mixed Laplace and $\Gamma(\cdot)$ distributions of the observation noise; on the right: resolution at $t = 0$ as a function of K^ℓ; the test resolution clearly exhibits $C\,K^{-1/3}$ behavior.

ϵ \ K	200	500	1000	2000	5000	10000	20000
1.0e-1	1.5e-2 1.82	1.5e-2 1.82	1.7e-2 1.78	1.6e-2 1.80	1.6e-2 1.80	1.6e-2 1.80	1.5e-2 1.82
1.0e-2	1.3e-3 1.45	1.2e-3 1.46	1.2e-3 1.46	1.2e-3 1.46	1.2e-3 1.46	1.2e-3 1.45	1.2e-3 1.46
1.0e-3	1.0e-4 1.33	0.9e-4 1.35	1.1e-4 1.32	1.1e-4 1.32	1.1e-4 1.32	0.9e-4 1.34	1.1e-4 1.32
1.0e-4	1.1e-5 1.24	0.9e-5 1.26	1.0e-5 1.25	0.9e-5 1.26	1.1e-5 1.24	0.7e-5 1.29	0.9e-5 1.26

Table 2: Quantifying conservatism of $\widehat{T}_\rho$ in Deconvolution experiment; in a cell: top – $\widehat{\epsilon}[K]$, bottom – the ratio $\frac{\ln \widehat{\epsilon}[K]}{\ln \epsilon[K]}$.

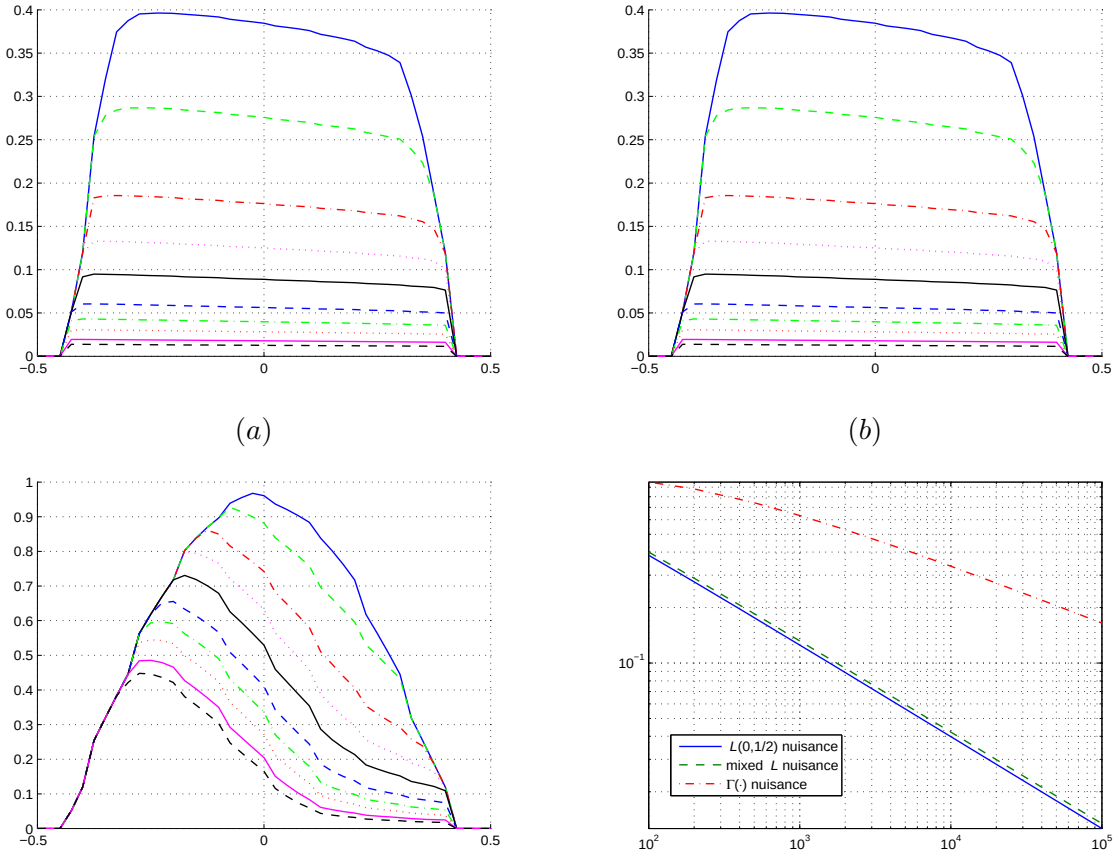

Figure 7: Trimmed observation experiment, resolution of the simple test for different K, $\epsilon = 0.05$, $\alpha = 0.5$; $K = [100, 200, 500, 1000, ..., 100\,000]$. Plot (a): resolution of the test as a function of $t \in [-1, 1]$, $L(0, \frac{1}{2})$ nuisance; plot (b) same for mixed Laplace nuisance; plot (c): resolution of the test with $\Gamma(\cdot)$ nuisance distribution. On plot (d): resolution at $t = 0$ as a function of sample size K. While the test resolution exhibits $C\,K^{-1/3}$ behavior in the case of Laplace an mixed Laplace nuisance, convergence is slow (if any) in the case of $\Gamma(\cdot)$ nuisance distribution.

4.4 Testing hypotheses on Markov chains

In this section, we present some applications of our approach to Markov chain related hypotheses testing. For a positive integer n, let $\Delta_n = \{x \in \mathbb{R}^n_+ : \sum_i x_i = 1\}$, and $\mathcal{S}_n$ be the set of all $n \times n$ stochastic matrices..

4.4.1 Deciding on two simple hypotheses

Situation. The simplest setting of the Markov chain related hypotheses testing is as follows. We are given two $n \times n$ stochastic matrices S^1 and S^2 with positive entries, specifying two hypotheses on an n-state Markov chain. Both hypotheses state that the probability distribution of the initial (at time 0) state ι_0 of the chain is a vector from some convex compact set $X \subset \mathrm{rint}\,\Delta_n$; in addition hypothesis H_1 (H_2) states that the transition matrix of the chain is S^1 (S^2). We observe on a given time horizon K a realization $\iota_0, \iota_1, ..., \iota_K$ of the trajectory of the chain and want to decide on the hypotheses.

Construction and result. With transition matrix fixed, the distribution of chain's trajectory on a fixed time horizon depends linearly on the distribution of the initial state. Consequently, our decision problem is to distinguish between two convex sets of probability distributions on the finite set of all possible chain trajectories from time 0 to time K inclusively. According to the Discrete case version of our results, a nearly optimal test is as follows: we solve the optimization problem

$$\varepsilon_\star = \max_{p,q\in X} \sum_{1\le\iota_0,\iota_1,...,\iota_K\le n} \sqrt{\left[p_{\iota_0} S^1_{\iota_1\iota_0} S^1_{\iota_2\iota_1} ... S^1_{\iota_K\iota_{K-1}}\right]\left[q_{\iota_0} S^2_{\iota_1\iota_0} S^2_{\iota_2\iota_1} ... S^2_{\iota_K\iota_{K-1}}\right]}; \tag{47}$$

denoting the optimal solution (p_*, q_*) and setting

$$\phi(\iota_0, ..., \iota_K) = \frac{1}{2} \ln\left(\frac{p_{\iota_0} S^1_{\iota_1\iota_0} S^1_{\iota_2\iota_1} ... S^1_{\iota_K\iota_{K-1}}}{q_{\iota_0} S^2_{\iota_1\iota_0} S^2_{\iota_2\iota_1} ... S^2_{\iota_K\iota_{K-1}}}\right),$$

the near-optimal test, the observed trajectory being $\iota^K = (\iota_0, ..., \iota_K)$, accepts H_1 when $\phi(\iota^K) \ge 0$, and accepts H_2 otherwise. The risk of this test is upper-bounded by $\varepsilon_\star$ given by (47).

Optimization problem (47) clearly is convex and solvable, and whenever (p, q) is feasible for the problem, so is (q, p), the values of the objective at these two solutions being the same. As a result, there exists an optimal solution (p_*, q_*) with $p_* = q_*$. The test ϕ associated with such a solution is completely independent of p_* and is just the plain likelihood ratio test:

$$\phi(\iota^K = (\iota_0, ..., \iota_K)) = \frac{1}{2} \sum_{\tau=1}^{K} \ln\left(\frac{S^1_{\iota_\tau\iota_{\tau-1}}}{S^2_{\iota_\tau\iota_{\tau-1}}}\right).$$

The (upper bound on the) risk of this test is immediately given by (47):

$$\varepsilon_\star = \max_{p\in X} \sum_{j=1}^{m} \left(\sum_{i=1}^{m} (S^1_{ij} S^2_{ij})^{t/2}\right) p_j.$$

$\lambda = 50$						$\lambda = 100$						$\lambda = 200$					
μ_1	μ_2	K	μ_1	μ_2	K	μ_1	μ_2	K	μ_1	μ_2	K	μ_1	μ_2	K	μ_1	μ_2	K
1.00	0.90	144	1.00	1.11	146	1.00	0.90	91	1.00	1.11	74	1.00	0.90	1929	1.00	1.11	1404
1.00	0.75	21	1.00	1.33	21	1.00	0.75	19	1.00	1.33	11	1.00	0.75	326	1.00	1.33	133
1.00	0.50	6	1.00	2.00	5	1.00	0.50	8	1.00	2.00	3	1.00	0.50	86	1.00	2.00	7

Table 3: Deciding with risk $\varepsilon_* = 0.01$ on two simple hypotheses on the parameter μ of a queuing system with $s = 100$, $b = 20$.

Numerical illustration. Consider a queuing system $(M/M/s/s+b)$ with s identical servers, with services times following exponential distribution $\mathcal{E}(\mu)$ with parameter μ, and a common buffer of capacity b. The input stream of customers is Poisson process with rate λ. Upon arrival, a customer either starts to be served, if there is a free server, or joins the buffer, if all servers are busy and there are less than b customers in the buffer, or leaves the system, if all servers are busy and there are b waiting customers in the buffer. The system is observed at time instances $0, 1, ..., K$, and we want to distinguish between two systems differing only in the value of μ, which is μ_1 for the first, and μ_2 for the second system. The observations form a Markov chain with $n = s + b + 1$ states, a state $j \in \{1, ..., n\}$ at time $t = 1, 2, ...$ meaning that at this time there are $s(j) := \min[j - 1, s]$ busy servers and $j - s(j) - 1$ customers in the buffer. Under hypothesis H_χ, $\chi = 1, 2$, the transition matrix of the chain is $S^\chi = \exp\{L^\chi\}$, where $L^\chi = L(\lambda, \mu_\chi)$ is a 3-diagonal *transition rate matrix* with zero column sums and $[L^\chi]_{j-1,j} = s(j)\mu_\chi$, $[L^\chi]_{j+1,j} = \lambda$. In table 3, we present a sample of (the smallest) observation times K ensuring that the upper bound $\varepsilon_\star$ on the risk of the simple test developed in this section is ≤ 0.01. We restrict ourselves to the case when distribution of the initial state is not subject to any restrictions, that is, $X = \Delta_{s+b+1}$.

4.4.2 Deciding on two composite hypotheses

In the previous example, we dealt with two simple hypotheses on a Markov chain with fully observable trajectory. Now consider the case of two composite hypotheses and indirect observations of state transitions.[14] More specifically, we intend to consider the case when a "composite hypothesis" specifies a set in $\mathcal{S}_n$ containing the transition matrix of the chain we are observing, and "indirectness of observations" means that instead of observing consecutive states of the chain trajectory, we are observing some encodings of these states (e.g., in the simplest case, the state space of the chain is split into non-overlapping subsets – *bins*, and our observations are the bins to which the consecutive states of the chain belong).

Preliminaries. Probability distribution P_t of the trajectories, on time horizon t, of a Markov chain depends nonlinearly on the transition matrix of the chain. As a result, to utilize our convexity-based approach, we need to work with composite hypotheses of "favorable structure," meaning that the family $\mathcal{P}_t$ of distributions P_t associated with transition matrices allowed by the hypothesis admits a reasonable convex approximation. We start with specifying the main ingredient of such "favorable structure."

[14] One problem of testing specific composite hypotheses about Markov chains has been studied in [11] using a closely related approach. The techniques we discuss here are different and clearly aimed at numerical treatment of the problem.

Let $K_1, ..., K_n$ be closed cones, all different from $\{0\}$, contained in $\mathbb{R}^n_+$. The collection $K^n = \{K_1, ..., K_n\}$ gives rise to the following two entities:

- The set of stochastic matrices

$$\mathcal{S} = \{S = [S_{ij}]_{i,j=1}^n \in \mathbb{R}^{n\times n} : \mathrm{Col}_j[S] \in K_j, \sum_i S_{ij} = 1,\ j = 1, ..., n\}$$

  (from now on, $\mathrm{Col}_j[S]$ is the j-th column of S);

- The convex set

$$\mathcal{P} = \{P = [P_{ij}]_{i,j=1}^n \in \mathbb{R}^{n\times n} : \mathrm{Col}_j[P] \in K_j,\ 1 \le j \le n, \sum_{i,j} P_{ij} = 1\}.$$

One has[15]

$$\mathcal{P} = \{P = [P_{ij}]_{i,j=1}^n : \exists (S \in \mathcal{S}, x \in \Delta_n) : \mathrm{Col}_j[P] = x_j \mathrm{Col}_j[S],\ j = 1, ..., n\}. \tag{48}$$

As a result, in a pair (S, x) associated with $P \in \mathcal{P}$ according to (48), x is uniquely defined by P:

$$x_j = \sum_i P_{ij},\ 1 \le j \le n;$$

besides this, for every j such that $\sum_i P_{ij} > 0$, $\mathrm{Col}_j[S]$ is the probabilistic normalization of $\mathrm{Col}_j[P]$.

Remark. The role played by the just defined entities in our context stems from the following immediate observation: consider a Markov chain with transition matrix S from $\mathcal{S}$, and let $x \in \Delta_n$ be the distribution of the state $\iota_{\tau-1}$ of this chain at time $\tau - 1$. Denoting by ι_τ the state of the chain at time τ, the distribution of the state transition $(\iota_{\tau-1}, \iota_\tau)$ clearly is

$$p_{ij} = S_{ij} x_j,\ 1 \le i, j \le n.$$

According to (48), $\mathcal{P}$ is nothing but the convex hull of all distributions of this type stemming from different $x \in \Delta_n$ and $S \in \mathcal{S}$.

Situation. Assume that for $\chi = 1, 2$ we are given

- collection of cones $K_\chi^{n_\chi} = \{K_1^\chi, ..., K_{n_\chi}^\chi\}$ of the type described in the preliminaries. This collection, as explained above, specifies a set $\mathcal{S}_\chi$ of stochastic $n_\chi \times n_\chi$ matrices and a set $\mathcal{P}_\chi$ of $n_\chi \times n_\chi$ matrices with nonnegative entries summing up to 1.
- $m \times n_\chi^2$ "observation matrix" A_χ with *positive entries* and unit column sums. We think of the n_χ^2 columns of A_χ as being indexed by the pairs (i, j), $1 \le i, j \le n_\chi$.

The outlined data specify, for $\chi = 1, 2$,

[15] Indeed, for $S \in \mathcal{S}$, $x \in \Delta_n$ the matrix P given by $\mathrm{Col}_j[P] = x_j \mathrm{Col}_j[S]$, $1 \le j \le n$, clearly belongs to $\mathcal{P}$. Vice versa, if $P \in \mathcal{P}$, then, setting $x_j = \sum_i P_{ij}$ and specifying the j-th column of S as $\mathrm{Col}_j[P]/x_j$ when $x_j \ne 0$ and as a whatever vector from $K_j \cap \Delta_n$ when $x_j = 0$, we get $S \in \mathcal{S}$, $x \in \Delta_n$ and $\mathrm{Col}_j[P] = x_j \mathrm{Col}_j[S]$ for all j.

- the family $\mathcal{M}_\chi$ of Markov chains. Chains from $\mathcal{M}_\chi$ have n_χ states, and their transition matrices belong to $\mathcal{S}_\chi$;
- observation scheme for transitions of a chain from $\mathcal{M}_\chi$. Specifically, observation ω_τ of the transition $\iota_{\tau-1} \to \iota_\tau$ takes values in $\{1, 2, ..., m\}$, and its conditional, the past of chain's state trajectory being given, distribution is the column $\mathrm{Col}_{(\iota_{\tau-1},\iota_\tau)}[A_\chi]$ of A_χ.

Now assume that "in the nature" there exist two Markov chains, indexed by $\chi = 1, 2$, with n_χ states and transition matrices S_χ, such that chain χ belongs to $\mathcal{M}_\chi$, and we observe one of these two chains as explained above, so that, independently of χ, our observation ω_t at time t takes values in $\{1, ..., m\}$. Given observation $\omega^K = (\omega_1, ..., \omega_K)$, we want to decide on the hypotheses H_χ, $\chi = 1, 2$, where H_χ states that the chain we are observing is chain χ.

Construction and result. We can approach our goal as follows. Every $P \in \mathcal{P}_\chi$ is a nonnegative $n_\chi \times n_\chi$ matrix with unit sum of entries and as such can be thought of as a probability distribution on $\mathcal{I}_\chi = \{(i, j) : 1 \le i, j \le n_\chi\}$. Matrix A_χ naturally associates with such a distribution a probability distribution $\mathcal{A}_\chi(P)$ on $\{1, ..., m\}$:

$$\mathcal{A}_\chi(P) = \sum_{i,j=1}^{n_\chi} P_{ij} \mathrm{Col}_{(i,j)}(A_\chi).$$

Note that the mapping $P \mapsto \mathcal{A}_\chi(P)$ is linear.

Let us define the convex compact subsets X^χ of the probabilistic simplex Δ_m by the relation

$$X^\chi = \{p \in \Delta_m : \exists P \in \mathcal{P}_\chi : p = \mathcal{A}_\chi(P)\}, \ \chi = 1, 2.$$

By the above remark,

> (!) *For a chain from $\mathcal{M}_\chi$ and every time instant $\tau \ge 1$, the conditional, given chain's trajectory prior to instant $\tau - 1$, distribution of the state transition $(\iota_{\tau-1}, \iota_\tau)$ belongs to $\mathcal{P}_\chi$, and, consequently, the conditional, by the same condition, distribution of the observation ω_τ belongs to X^χ.*

Note that $X^\chi \subset \mathrm{rint}\, \Delta_m$ due to entrywise positivity of A_χ.

For $t = 1, 2, ...$, let $\jmath_{t,1}$, $\jmath_{t,2}$ be the states of chain 1 and chain 2 at time t, let $\zeta_{t,\chi} = (\jmath_{t,\chi}, \jmath_{t-1,\chi})$, $\chi = 1, 2$, and let $X_t = X^1$, $Y_t = X^2$. With this setup, we arrive at the situation considered in Proposition 3.2: for $\chi = 1, 2$, under hypothesis H_χ ω_t is a deterministic function of $\zeta_\chi^t = (\zeta_{1,\chi}, ..., \zeta_{t,\chi})$, the conditional, given ζ_χ^{t-1}, distribution of ω_t depends deterministically on ζ_χ^{t-1} and, by (!), belongs to X^χ. Hence, Proposition 3.2 implies

Proposition 4.4 *In the situation and under assumptions of this section, let the sets X^1, X^2 do not intersect. Let p_1^*, p_2^*, form the optimal solution to the problem*

$$\varepsilon_\star = \max_{p_1, p_2} \left\{ \sum_{\omega=1}^m \sqrt{[p_1]_\omega [p_2]_\omega} : \ p_1 \in X^1, \ p_2 \in X^2, \right\}, \tag{49}$$

and let

$$\phi(\omega) = \frac{1}{2} \ln \left(\frac{[p_1^*]_\omega}{[p_2^*]_\omega} \right).$$

Then the risk of the test which, given observations $\omega_1, ..., \omega_K$, accepts H_2 when $\sum_{\tau=1}^K \phi(\omega_\tau) \ge 0$ and accepts H_2 otherwise, is at most $\varepsilon_\star^K$.

Remark. By inspecting the proof, Proposition 4.4 remains valid in the situation where $\mathcal{M}_\chi$ are families of *non-stationary* Markov chains with n_χ states $1, ..., n_\chi$. In such a chain, for every $\tau > 0$, the conditional, given the trajectory $\iota_0, ..., \iota_{\tau-1}$ of the chain from time 0 to time $\tau - 1$, distribution of state ι_τ at time τ is selected, in a non-anticipative fashion, from the set $K^\chi_{\iota_{\tau-1}} \cap \Delta_n$.

Numerical illustration: random walk. Consider a toy example where the Markov chains $\mathcal{M}_\chi$, $\chi = 1, 2$, represent a random walk along $n = 16$-element grid on the unit circle; thus, each chain has 16 states. The "nominal" transition matrices S^n_χ correspond to the walk where one stays in the current position with probability $1 - 2p_\chi$ and jumps to a neighbouring position with probability $2p_\chi$, with equal probabilities to move clock- and counter-clockwise; in our experiment, $p_1 = 0.2$ and $p_2 = 0.4$. The actual transition matrix S_χ of chain $\mathcal{M}_\chi$ is allowed to belong to the "uncertainty set"

$$\mathcal{U}_\chi = \{S_\chi \in \mathcal{S}_n : (1-\rho)S^n_\chi \le S_\chi \le (1+\rho)S^n_\chi\},$$

where the inequalities are entrywise. In other words, the cones K^χ_j, $j = 1, 2, ..., n$, are the conic hulls of the sets

$$\{q \in \Delta_n : (1-\rho)\mathrm{Col}_j[S^n_\chi] \le q \le (1+\rho)\mathrm{Col}_j[S^n_\chi]\}.$$

In our experiments, we used $\rho = 0.1$.

We have considered two observation schemes: "direct observations", where we observe the positions of the walker at times 0,1,..., and "indirect observations," where the 16 potential positions are split into 8 "bins," two states per bin, and what we see at time instant t is the bin to which t-th position of the walker belongs. In the latter case we used a random partition of the states into the bins which was common for the chains $\mathcal{M}_1$ and $\mathcal{M}_2$ (i.e., in our experiments the "observation matrices" A_1 and A_2 always coincided with each other).

The results of a typical experiment are presented in table 4. For each of our two observation schemes, we start with observation time which, according to Proposition 4.4, guarantees the risk $\epsilon = 0.01$, and then decrease the observation time to see how the performance of the test deteriorates. In different simulations, we used different transition matrices allowed by the corresponding hypotheses, including the "critical" ones – those associated with the optimal solution to (49). Evaluating the results of the experiment is not easy – in the first place, it is unclear what could be a natural "benchmark" to be compared to, especially when the observations are indirect. In the case of direct observations we have considered as a contender the likelihood ratio test (see section 4.4.1) straightforwardly adjusted to the uncertainty in the transition matrix.[16] Such test turns out to be essentially less precise than the test presented in Proposition 4.4; e.g., in the experiment reported in column A of table 4, with observation time 71 the risks of the adjusted likelihood test were as large as 0.01/0.06.

4.4.3 Two composite hypotheses revisited

In the situation of section 4.4.2 (perhaps, indirect) observations of *transitions* of a Markov chain were available. We are about to consider the model in which we are only allowed to observe how

[16] Specifically, given the chain trajectory $\iota_0, ..., \iota_t$, we can easily compute the maximal and the minimal values, $\psi_{\max}$ and $\psi_{\min}$, of the logarithm of likelihood ratio as allowed by our uncertainties in the transition matrices. Namely, $\psi_{\max} = \max_{\{S_{\tau,1}, S_{\tau,2}\}_{\tau=1}^t} \sum_{\tau=1}^t \ln([S_{\tau,1}]_{\jmath_\tau,\jmath_{\tau-1}}/[S_{\tau,2}]_{\jmath_\tau,\jmath_{\tau-1}})$, where $S_{\tau,\chi}$ run through the uncertainty sets associated with hypotheses H_χ, $\chi = 1, 2$; $\psi_{\min}$ is defined similarly, with $\max_{\{S_{\tau,1}, S_{\tau,2}\}_{\tau=1}^t}$ replaced with $\min_{\{S_{\tau,1}, S_{\tau,2}\}_{\tau=1}^t}$. We accept H_1 when a randomly selected point in $[\psi_{\min}, \psi_{\max}]$ turns out to be nonnegative, and accept H_2 otherwise.

(a) $\varepsilon_\star = 0.9368$

t	$\varepsilon_\star^t$	Risk(T)	Risk(ML)
71	0.0097	0.0004/0.0008	0.0094/0.0551
48	0.0436	0.0038/0.0018	0.0192/0.0798
32	0.1239	0.0226/0.0118	0.0390/0.1426
21	0.2540	0.0230/0.0610	0.0620/0.1903
14	0.4011	0.0870/0.0508	0.1008/0.2470
10	0.5207	0.0780/0.1412	0.1268/0.2649
7	0.6333	0.1184/0.1688	0.1824/0.3368
5	0.7216	0.1040/0.2682	0.2190/0.2792
3	0.8222	0.3780/0.1166	0.3000/0.4027
2	0.8777	0.1814/0.3780	0.1814/0.3780
1	0.9368	0.4230/0.2064	0.4230/0.2064

(b) $\varepsilon_\star = 0.9880$

t	$\varepsilon_\star^t$	Risk(T)
381	0.0099	0.0000/0.0000
254	0.0462	0.0000/0.0000
170	0.1277	0.0000/0.0002
113	0.2546	0.0002/0.0008
76	0.3982	0.0002/0.0054
51	0.5393	0.0022/0.0168
34	0.6626	0.0086/0.0412
23	0.7569	0.0210/0.0758
15	0.8339	0.0540/0.1018
10	0.8860	0.0872/0.1530
7	0.9187	0.1420/0.1790
5	0.9413	0.1386/0.2878
3	0.9643	0.2812/0.2638
2	0.9761	0.2078/0.3824
1	0.9880	0.3816/0.2546

Table 4: Random walk. (a) - direct observations; (b) - indirect observations. In the table: t: observation time; $\varepsilon_\star^t$ and Risk(T): theoretical upper bound on the risk of the test from Proposition 4.4, and empirical risk of the test; Risk(ML): empirical risk of the likelihood ratio test adjusted for uncertainty in transition probabilities. ϵ_1/ϵ_2 in "risk" columns: empirical, over 5000 simulations, probabilities to reject hypothesis H_1 (ϵ_1) and H_2 (ϵ_2) when the hypothesis is true. Partition of 16 states of the walk into 8 bins in the reported experiment is $\{1,8\}$, $\{4,6\}$, $\{5,7\}$, $\{9,11\}$, $\{3,19\}$, $\{2,15\}$, $\{12,16\}$, $\{13,14\}$.

frequently the chain visited different (groups of) states on a given time horizon, but do not use information in which order these states were visited.

Preliminaries. For $Q \in \mathcal{S}_n$ and $\rho \geq 0$, let

$$\mathcal{S}_n(Q,\rho) = \{S \in \mathcal{S}_n : \|S - Q\|_{1,1} \leq \rho\},$$

where for a $p \times q$ matrix C

$$\|C\|_{1,1} = \max_{1\leq j\leq q} \|\mathrm{Col}_j[C]\|_1$$

is the norm of the mapping $u \mapsto Cu : \mathbb{R}^q \times \mathbb{R}^p$ induced by the norms $\|\cdot\|_1$ on the argument and the image spaces.

Situation we consider here is as follows. "In the nature" there exist two Markov chains, indexed by $\chi = 1, 2$. Chain χ has n_χ states and transition matrix S_χ. Same as in section 4.4.2, we do not observe the states exactly, and our observation scheme is as follows. For $\chi = 1, 2$, we are given $m \times n_\chi$ matrices A_χ with positive entries and all column sums equal to 1. When observing chain χ, our observation η_τ at time τ takes values $1, ..., m$, and the conditional, given

the trajectory of the chain since time 0 to time τ inclusively, distribution of η_τ is the ι_τ-th column $\mathrm{Col}_{\iota_\tau}[A_\chi]$ of A_χ.

Now assume that all we know about S_χ, $\chi = 1, 2$, is that $S_\chi \in \mathcal{S}_{n_\chi}(Q_\chi, \rho_\chi)$ with known Q_χ and ρ_χ. We observe the sequence $\eta^t = (\eta_1, ..., \eta_t)$ coming from one of two chains, and want to decide on the hypotheses H_χ, $\chi = 1, 2$, stating that $S_\chi \in \mathcal{S}_{n_\chi}(Q_\chi, \rho_\chi)$.

Construction and result. Our approach is as follows. Given a positive integer κ, for $\chi = 1, 2$ let

$$Z_\chi = \mathrm{Conv}\{A_\chi v : v \in \Delta_{n_\chi}, \text{ and } \exists j : \|v - \mathrm{Col}_j[Q_\chi^\kappa]\|_1 \le \kappa\rho_\chi\} \subset \Delta_m.$$

Note that $Z_\chi \subset \mathrm{rint}\,\Delta_m$ (since the column sums in A_χ are equal to one, and all entries of A_χ are positive).

It is immediately seen that

- Under hypothesis H_χ, $\chi = 1, 2$, for every positive integer t, the conditional, given the state $\jmath_{\kappa(t-1),\chi}$ of the Markov chain χ at time $\kappa(t-1)$, distribution of observation $\eta_{\kappa t}$ belongs to Z_χ.

  Indeed, S_χ and Q_χ are stochastic matrices with $\|S_\chi - Q_\chi\|_{1,1} \le \rho_\chi$ (we are under hypothesis H_χ), and for stochastic matrices $A, B, \bar{A}$ and $\bar{B}$ one has

  $$\|\bar{A}\bar{B} - AB\|_{1,1} \le \|\bar{A} - A\|_{1,1} + \|\bar{B} - B\|_{1,1}$$

  due to

  $$\begin{array}{l}\|\bar{A}\bar{B} - AB\|_{1,1} \le \|\bar{A}(\bar{B} - B)\|_{1,1} + \|(\bar{A} - A)B\|_{1,1}\\ \le \|\bar{A}\|_{1,1}\|\bar{B} - B\|_{1,1} + \|\bar{A} - A\|_{1,1}\|B\|_{1,1} = \|\bar{B} - B\|_{1,1} + \|\bar{A} - A\|_{1,1}.\end{array}$$

  Whence $\|S_\chi^\kappa - Q_\chi^\kappa\|_{1,1} \le \kappa\rho_\chi$, so that the probabilistic vector $v = \mathrm{Col}_{\jmath_{\kappa(t-1),\chi}}[S_\chi^\kappa]$ satisfy $\|v - \mathrm{Col}_{\jmath_{\kappa(t-1),\chi}}[Q_\chi^\kappa]\|_1 \le \kappa\rho_\chi$. We conclude that the distribution of $A_\chi v$ of $\eta_{\kappa t}$ belongs to Z_χ.

- Z_χ is a polyhedral convex set with an explicit representation:

$$Z_\chi = \left\{ z : \exists \alpha, v^1, ..., v^{n_\chi} \in \mathbb{R}^{n_\chi} : \begin{array}{l} z = A_\chi \sum_{j=1}^{n_\chi} v^j,\ v^j \ge 0,\ \sum_{i=1}^{n_\chi} v_i^j = \alpha_j,\ \alpha \in \Delta_{n_\chi},\\ \|v^j - \alpha_j \mathrm{Col}_j[Q_\chi^\kappa]\|_1 \le \alpha_j\kappa\rho_\chi,\ \ 1 \le j \le n_\chi. \end{array} \right\}$$

Setting $\omega_t = \eta_{\kappa t}$, $\zeta_{t,\chi} = \jmath_{t\kappa,\chi}$, $\chi = 1, 2$, and $X_t = Z_1$, $Y_t = Z_2$, $t = 1, 2, ...$, we arrive at the situation considered in Proposition 3.2: under hypothesis H_χ, $\chi = 1, 2$, ω_t is a deterministic function of $\zeta_\chi^t = (\zeta_{0,\chi}, ..., \zeta_{t,\chi})$, and the conditional, given ζ_χ^{t-1}, distribution of ω_t is $\mu_t = A_\chi \mathrm{Col}_{\jmath_{(t-1)\kappa,\chi}}[S_\chi^\kappa]$, which is a deterministic function of ζ_χ^{t-1}. Besides this, $\mu_t \in X_t \equiv Z_1$ under hypothesis H_1, and $\mu_t \in Y_t \equiv Z_2$ under hypothesis H_2. For these reasons, Proposition 3.2 implies

Proposition 4.5 *Let κ be such that Z_1 does not intersect Z_2. Let, further, (x_*, y_*) be an optimal solution to the convex optimization problem*

$$\varepsilon_\star = \max_{x \in Z_1, y \in Z_2} \sum_{i=1}^m \sqrt{x_i y_i},$$

and let

$$\phi_*(i) = \frac{1}{2}\ln([x_*]_i/[y_*]_i),\ 1 \le i \le m.$$

Then for every positive integer K, the risk of the test ϕ_*^K which, given observation ω^K, accepts H_1 whenever

$$\sum_{t=1}^{K}\phi_*(\omega_t)=\sum_{i=1}^{m}\phi_*(i)\mathrm{Card}\{t\le K:\omega_t=i\} \tag{50}$$

is nonnegative and accepts H_2 otherwise, does not exceed $\varepsilon_\star^K$.

Remarks. Note that κ meeting the premise of Proposition 4.5 does exist, provided that ρ_χ are small enough and that $A_1e\neq A_2f$ for every pair of steady-state distributions $e=Q_1e$, $f=Q_2f$ of the chains with transition matrices Q_1 and Q_2.

Note that in order to compute the test statistics (50) *we do not need to observe the trajectory $\omega_1,\omega_2,...,\omega_K$; all what matters is the "histogram" $\{p_i=\mathrm{Card}\{t\le K:\omega_t=i\}\}_{i=1}^m$ of $\omega_1,...,\omega_K$.* Furthermore, we lose nothing if instead of observing *a single and long* ω-trajectory, we observe a *population of independent "short" trajectories.* Indeed, assume that N independent trajectories are observed on time horizon $L\kappa\le K\kappa$; all the trajectories start at time $\tau=0$ in a once for ever fixed state and then move from state to state independently of each other and utilizing the same transition matrix S. Our observations now are the total, over N trajectories, numbers p_i, $i=1,...,m$, of time instants of the form κt, $t\ge1$, spent by the trajectories in state i. If our goal is to decide which of the chains $\chi=1,2$ we are observing, it is immediately seen that Proposition 3.2 implies that under the premise and in the notation of Proposition 4.5, the test which accepts H_1 when $\sum_{i=1}^m\phi_*(i)p_i\ge0$ and accepts H_2 otherwise (cf. (50)) obeys the upper risk bound $\varepsilon_\star^{LN}$. In other words, the risk of the test would be exactly the same as if instead of (aggregated partial) information on N trajectories of length $L\kappa$ each we were collecting similar information on a single trajectory of length $K=LN\kappa$.

Numerical illustration. Consider a queuing system $(M/M/s/s+b)$ with several identical servers and a single buffer of capacity b. The service times of each server and inter-arrival times are exponentially distributed, with distributions $\mathcal{E}(\mu)$ and $\mathcal{E}(\lambda)$ respectively. Upon arrival, a customer either starts being served, when there are free servers, or joins the buffer queue, if all servers are busy and there are $<b$ customers in the buffer queue, or leaves the system immediately when all servers are busy and there are b customers in the buffer. We assume that the parameters λ, μ are not known exactly; all we know is that

$$|\lambda-\bar\lambda|\le\delta_\lambda\text{ and }|\mu-\bar\mu|\le\delta_\mu,$$

with given $\bar\lambda>0$, $\bar\mu>0$ and $\delta_\lambda<\bar\lambda$, $\delta_\mu<\bar\mu$.

We observe the number of customers in the buffer at times $t=1,2,...$, and want to decide on the hypotheses H_1 stating that the number of servers in the system is s_1, and H_2, stating that this number is s_2.

In terms of the hidden Markov chain framework presented above, the situation is as follows. Under hypothesis H_χ the queuing system can be modeled by Markov chain with $n_\chi=s_\chi+b+1$ states with the transition matrix of the chain $S_\chi=\exp\{L_\chi\}$, where the transition rate matrix $L_\chi=L_\chi(\lambda,\mu)$ satisfies

$$[L_\chi]_{j-1,j}=s(j)\mu,\ \ [L_\chi]_{j,j}=-(s(j)\mu+\lambda),\ \ [L_\chi]_{j+1,j}=\lambda,\ \ s(j):=\min[j-1,s_\chi],\ 1\le j\le n_\chi.$$

It is immediately seen that if $Q_\chi = \exp\{L_\chi(\bar\lambda, \bar\mu)\}$, it holds[17]

$$\|S_\chi - Q_\chi\|_{1,1} \le \rho_\chi := 2\delta_\lambda + 2s_\chi\delta_\mu.$$

We can now apply the outlined scheme to decide between the hypotheses H_1 and H_2. A numerical illustration is presented in table 5; in this illustration, we use $\kappa = 1$, that is, observations used in the test are the numbers of customers in the buffer at times $t = 1, 2, ..., K$.

			$K = K_*$		$K = \lfloor K_*/2 \rfloor$		$K = \lfloor K_*/3 \rfloor$	
s_1, s_2, b	$\varepsilon_\star$	K_*	ϵ_1	ϵ_2	ϵ_1	ϵ_2	ϵ_1	ϵ_2
$s_1 = 10,\ s_2 = 9,\ b = 5$	0.993240	679	0.0000	0.0000	0.0035	0.0015	0.0119	0.0104
$s_1 = 10,\ s_2 = 7,\ b = 5$	0.894036	42	0.0002	0.0002	0.0093	0.0100	0.0260	0.0273

Table 5: Experiments with toy queuing systems. $\bar\lambda = 40, \bar\mu = 5, \rho_1 = \rho_2 = 0$. ϵ_χ: empirical, over sample of 10^4 experiments with observation time K each, probability to reject H_χ when the hypothesis is true. $\varepsilon_\star$ is defined in Proposition 4.5, $K_* = \lfloor \ln(1/0.01)/\ln(1/\varepsilon_\star)\rfloor$ is the observation time, as defined by Proposition 4.5, resulting in risk ≤ 0.01.

References

[1] E. D. Andersen and K. D. Andersen. *The MOSEK optimization toolbox for MATLAB manual. Version 7.0*, 2013. http://docs.mosek.com/7.0/toolbox/.

[2] A. Antoniadis and I. Gijbels. Detecting abrupt changes by wavelet methods. *Journal of Nonparametric Statistics*, 14(1-2):7–29, 2002.

[3] T. Augustin and R. Hable. On the impact of robust statistics on imprecise probability models: a review. *Structural Safety*, 32(6):358–365, 2010.

[4] M. Basseville. Detecting changes in signals and systems – a survey. *Automatica*, 24(3):309–326, 1988.

[5] T. Bednarski et al. Binary experiments, minimax tests and 2-alternating capacities. *The Annals of Statistics*, 10(1):226–232, 1982.

[6] A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, volume 2. Siam, 2001.

[7] L. Birgé. *Approximation dans les spaces métriques et théorie de l'estimation: inégalités de Cràmer-Chernoff et théorie asymptotique des tests*. PhD thesis, Université Paris VII, 1980.

[8] L. Birgé. Vitesses maximales de décroissance des erreurs et tests optimaux associés. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 55(3):261–273, 1981.

[17] Indeed, we have $S_\chi = \lim_{k\to\infty}(I + \frac{1}{k}L_\chi(\lambda,\mu))^k$; for large k, the matrix $N_k(\lambda,\chi) = I + \frac{1}{k}L_\chi(\lambda,\mu)$ is stochastic, and we clearly have $\|N_k(\lambda,\mu) - N_k(\bar\lambda,\bar\mu)\|_{1,1} \le k^{-1}\rho_\chi$. Whence, as we have already seen,

$$\|N_k^k(\lambda,\mu) - N_k^k(\bar\lambda,\bar\mu)\|_{1,1} \le \rho_\chi.$$

When passing to the limit as $k \to \infty$, we get the desired bound on $\|S_\chi - Q_\chi\|_{1,1}$.

[9] L. Birgé. Sur un théorème de minimax et son application aux tests. *Probab. Math. Stat.*, 3:259–282, 1982.

[10] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 65(2):181–237, 1983.

[11] L. Birgé. Robust testing for independent non identically distributed variables and Markov chains. In *Specifying Statistical Models*, pages 134–162. Springer, 1983.

[12] L. Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 42, pages 273–325. Elsevier, 2006.

[13] A. Buja. On the huber-strassen theorem. *Probability Theory and Related Fields*, 73(1):149–152, 1986.

[14] M. Burnashev. On the minimax detection of an imperfectly known signal in a white noise background. *Theory Probab. Appl.*, 24:107–119, 1979.

[15] M. Burnashev. Discrimination of hypotheses for gaussian measures and a geometric characterization of the gaussian distribution. *Math. Notes*, 32:757–761, 1982.

[16] I. Dattner, A. Goldenshluger, A. Juditsky, et al. On deconvolution of distribution functions. *The Annals of Statistics*, 39(5):2477–2501, 2011.

[17] D. Donoho. Statistical estimation and optimal recovery. *The Annals of Statistics*, 22(1):238–270, 1994.

[18] D. Donoho and R. Liu. Geometrizing rate of convergence I. Technical report, Tech. Report 137a, Dept. of Statist., University of California, Berkeley, 1987.

[19] D. L. Donoho and R. C. Liu. Geometrizing rates of convergence, II. *The Annals of Statistics*, pages 633–667, 1991.

[20] J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, pages 1257–1272, 1991.

[21] W. R. Gaffey et al. A consistent estimator of a component of a convolution. *The Annals of Mathematical Statistics*, 30(1):198–205, 1959.

[22] A. Goldenshluger, A. Juditsky, A. Tsybakov, and A. Zeevi. Change–point estimation from indirect observations. 1. minimax complexity. *Ann. Inst. Henri Poincare Probab. Stat.*, 44:787–818, 2008.

[23] A. Goldenshluger, A. Juditsky, A. Tsybakov, and A. Zeevi. Change-point estimation from indirect observations. 2. adaptation. *Ann. Inst. H. Poincare Probab. Statist*, 44(5):819–836, 2008.

[24] M. Grant and S. Boyd. *The* `CVX` *Users Guide. Release 2.1*, 2014. http://web.cvxr.com/cvx/doc/CVX.pdf.

[25] F. Gustafsson. *Adaptive filtering and change detection*, volume 1. Wiley New York, 2000.

[26] P. J. Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, 36(6):1753–1758, 1965.

[27] P. J. Huber and V. Strassen. Minimax tests and the Neyman-Pearson lemma for capacities. *The Annals of Statistics*, 1(2):251–263, 1973.

[28] P. J. Huber, V. Strassen, et al. Note: Correction to minimax tests and the Neyman-Pearson lemma for capacities. *The Annals of Statistics*, 2(1):223–224, 1974.

[29] I. A. Ibragimov and R. Z. Khas' minskii. On nonparametric estimation of the value of a linear functional in gaussian white noise. *Theory of Probability & Its Applications*, 29(1):18–32, 1985.

[30] I. A. Ibragimov and R. Z. Khas' minskii. Estimation of linear functionals in gaussian noise. *Theory of Probability & Its Applications*, 32(1):30–39, 1988.

[31] Y. Ingster and I. A. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169 of *Lecture Notes in Statistics*. Springer, 2002.

[32] A. B. Juditsky and A. S. Nemirovski. Nonparametric estimation by convex programming. *The Annals of Statistics*, 37(5a):2278–2300, 2009.

[33] V. Kuznetsov. Stable detection when signal and spectrum of normal noise are inaccurately known. *Telecommunications and radio engineering*, 30(3):58–64, 1976.

[34] L. Le Cam. On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *The Annals of Mathematical Statistics*, pages 802–828, 1970.

[35] L. Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.

[36] L. Le Cam. On local and global properties in the theory of asymptotic normality of experiments. *Stochastic processes and related topics*, 1:13–54, 1975.

[37] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. Springer, 1986.

[38] A. Meister. *Deconvolution problems in nonparametric statistics*, volume 193. Springer, 2009.

[39] H.-G. Müller and U. Stadtmüller. Discontinuous versus smooth regression. *The Annals of Statistics*, 27(1):299–337, 1999.

[40] M. H. Neumann. Optimal change-point estimation in inverse problems. *Scandinavian Journal of Statistics*, 24(4):503–521, 1997.

[41] F. Österreicher. On the construction of least favourable pairs of distributions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 43(1):49–55, 1978.

[42] H. Rieder. Least favorable pairs for special capacities. *The Annals of Statistics*, pages 909–921, 1977.

[43] A. G. Tartakovsky and V. V. Veeravalli. Change-point detection in multichannel and distributed systems. *Applied Sequential Methodologies: Real-World Examples with Data Analysis*, 173:339–370, 2004.

[44] A. G. Tartakovsky and V. V. Veeravalli. Asymptotically optimal quickest change detection in distributed sensor systems. *Sequential Analysis*, 27(4):441–475, 2008.

[45] Y. Wang. Jump and sharp cusp detection by wavelets. *Biometrika*, 82(2):385–397, 1995.

[46] Y. Yin. Detection of the number, locations and magnitudes of jumps. *Communications in Statistics. Stochastic Models*, 4(3):445–455, 1988.

[47] C.-H. Zhang. Fourier methods for estimating mixing densities and distributions. *The Annals of Statistics*, pages 806–831, 1990.

A Proofs

A.1 Proof of Theorem 2.1

1^0. The fact that the function (2) is continuous on its domain, convex in $\phi(\cdot) \in \mathcal{F}$ and concave in $[x; y] \in X \times Y$ is readily given by our basic assumptions. Let us set

$$\Psi([x;y]) = \inf_{\phi\in\mathcal{F}} \Phi(\phi, [x;y]). \tag{51}$$

We claim that the function

$$\phi_{x,y}(\omega) = \frac{1}{2}\ln(p_x(\omega)/p_y(\omega))$$

(which, by our assumptions, belongs to $\mathcal{F}$) is an optimal solution to the right hand side minimization problem in (51), so that

$$\forall(x \in X, y \in Y) : \Psi([x;y]) := \inf_{\phi\in\mathcal{F}} \Phi(\phi, [x;y]) = 2\ln\left(\int_\Omega \sqrt{p_x(\omega)p_y(\omega)}P(d\omega)\right). \tag{52}$$

Note that Ψ, being the infinum of a family of concave functions of $[x; y] \in \mathcal{M} \times \mathcal{M}$, is concave on $\mathcal{M} \times \mathcal{M}$. Indeed, we have

$$\exp\{-\phi_{x,y}(\omega)\}p_x(\omega) = \exp\{\phi_{x,y}(\omega)\}p_y(\omega) = g(\omega) := \sqrt{p_x(\omega)p_y(\omega)},$$

whence $\Phi(\phi_{x,y}, [x; y]) = 2\ln\left(\int_\Omega g(\omega)P(d\omega)\right)$. On the other hand, for $\phi(\cdot) = \phi_{x,y}(\cdot) + \delta(\cdot) \in \mathcal{F}$ we have

$$\begin{array}{ll} & \int_\Omega g(\omega)P(d\omega) = \int_\Omega \left[\sqrt{g(\omega)}\exp\{-\delta(\omega)/2\}\right]\left[\sqrt{g(\omega)}\exp\{\delta(\omega)/2\}\right]P(d\omega) \\ (a) & \le \left(\int_\Omega g(\omega)\exp\{-\delta(\omega)\}P(d\omega)\right)^{1/2}\left(\int_\Omega g(\omega)\exp\{\delta(\omega)\}P(d\omega)\right)^{1/2} \\ & = \left(\int_\Omega \exp\{-\phi(\omega)\}p_x(\omega)P(d\omega)\right)^{1/2}\left(\int_\Omega \exp\{\phi(\omega)\}p_y(\omega)P(d\omega)\right)^{1/2} \\ (b) & \Rightarrow 2\ln\left(\int_\Omega g(\omega)P(d\omega)\right) \le \Phi(\phi, [x;y]), \end{array}$$

and thus $\Phi(\phi_{x,y}, [x, y]) \le \Phi(\phi, [x; y])$ for every $\phi \in \mathcal{F}$.

Remark A.1 Note that the inequality in (b) can be equality only when the inequality in (a) is so. In other words, if $\bar{\phi}$ is a minimizer of $\Phi(\phi,[x;y])$ over $\phi\in\mathcal{F}$, setting $\delta(\cdot)=\bar{\phi}(\cdot)-\phi_{x,y}(\cdot)$, the functions $\sqrt{g(\omega)}\exp\{-\delta(\omega)/2\}$ and $\sqrt{g(\omega)}\exp\{\delta(\omega)/2\}$, considered as elements of $L_2[\Omega,P]$, are proportional to each other. Since g is positive and g,δ are continuous, while the support of P is the entire Ω, this "L_2-proportionality" means that the functions in question differ by a constant factor, or, which is the same, that $\delta(\cdot)$ is constant. Thus, *the minimizers of* $\Phi(\phi,[x;y])$ *over* $\phi\in\mathcal{F}$ *are exactly the functions of the form* $\phi(\omega)=\phi_{x,y}(\omega)+\text{const}$.

2^0. We are about to verify that $\Phi(\phi,[x;y])$ has a saddle point (min in $\phi\in\mathcal{F}$, max in $[x;y]$) on $\mathcal{F}\times(X\times Y)$. Indeed, observe, first, that on the domain of Φ it holds

$$\Phi(\phi(\cdot)+a,[x;y])=\Phi(\phi(\cdot),[x;y])\ \forall(a\in\mathbb{R},\phi\in\mathcal{F}). \tag{53}$$

Thus, it suffices to verify that $\Phi(\phi,[x;y])$ has a saddle point on the set $\mathcal{F}_0\times(X\times Y)$, with $\mathcal{F}_0=\{\phi\in\mathcal{F}:\int_\Omega\phi(\omega)P(d\omega)=0\}$. Since $X\times Y$ is a convex compact set, Φ is continuous on $\mathcal{F}_0\times(X\times Y)$ and convex-concave, all we need in order to verify the existence of a saddle point is to show that Φ is coercive in the first argument, that is, for every fixed $[x;y]\in X\times Y$ one has $\Phi(\phi,[x;y])\to+\infty$ as $\|\phi\|\to\infty$ (whatever be the norm $\|\cdot\|$ on $\mathcal{F}_0$; recall that $\mathcal{F}_0$ is a finite-dimensional linear space). Setting $\Theta(\phi)=\Phi(\phi,[x;y])$ and taking into account that Θ is convex and finite on $\mathcal{F}_0$, in order to prove that Θ is coercive, it suffices to verify that $\Theta(t\phi)\to\infty$, $t\to\infty$, for every nonzero $\phi\in\mathcal{F}_0$, which is evident: since $\int_\Omega\phi(\omega)P(d\omega)=0$ and ϕ is nonzero, we have $\int_\Omega\max[\phi(\omega),0]P(d\omega)=\int_\Omega\max[-\phi(\omega),0]P(d\omega)>0$, whence $\Theta(t\phi)\to\infty$ as $t\to\infty$ due to the fact that both $p_x(\cdot)$ and $p_y(\cdot)$ are positive everywhere.

3^0. Now let $(\phi_*(\cdot);[x_*;y_*])$ be a saddle point of Φ on $\mathcal{F}\times(X\times Y)$. Shifting, if necessary, $\phi_*(\cdot)$ by a constant (by (53), this does not affect the fact that $(\phi_*,[x_*;y_*])$ is a saddle point of Φ), we can assume that

$$\varepsilon_\star:=\int_\Omega\exp\{-\phi_*(\omega)\}p_{x_*}(\omega)P(d\omega)=\int_\Omega\exp\{\phi_*(\omega)\}p_{y_*}(\omega)P(d\omega), \tag{54}$$

so that the saddle point value of Φ is

$$\Phi_*:=\max_{[x;y]\in X\times Y}\min_{\phi\in\mathcal{F}}\Phi(\phi,[x;y])=\Phi(\phi_*,[x_*;y_*])=2\ln(\varepsilon_\star). \tag{55}$$

The following lemma completes the proof of Theorem 2.1.i:

Lemma A.1 *Under the premise of Theorem 2.1, let* $(\phi_*,[x_*;y_*])$ *be a saddle point of* Φ *satisfying* (54)*, and let* $\phi_*^a(\cdot)=\phi_*(\cdot)-a$, $a\in\mathbb{R}$. *Then*

$$\begin{array}{lrcl}(a) & \int_\Omega\exp\{-\phi_*^a(\omega)\}p_x(\omega)P(d\omega) & \le & \exp\{a\}\varepsilon_\star\ \forall x\in X,\\ (b) & \int_\Omega\exp\{\phi_*^a(\omega)\}p_y(\omega)P(d\omega) & \le & \exp\{-a\}\varepsilon_\star\ \forall y\in Y.\end{array} \tag{56}$$

As a result, for the simple test associated with the detector ϕ_*^a, *the probabilities* ϵ_X *to reject* H_X *when the hypothesis is true and* ϵ_Y *to reject* H_Y *when the hypothesis is true can be upper-bounded according to* (4).

Proof. For $x \in X$, we have

$$\begin{array}{rcl} 2\ln(\varepsilon_\star) & = & \Phi_* \geq \Phi(\phi_*, [x; y_*]) \\ & = & \ln\left(\int_\Omega \exp\{-\phi_*(\omega)\} p_x(\omega) P(d\omega)\right) + \ln\left(\int_\Omega \exp\{\phi_*(\omega)\} p_{y_*}(\omega) P(d\omega)\right) \\ & = & \ln\left(\int_\Omega \exp\{-\phi_*(\omega)\} p_x(\omega) P(d\omega)\right) + \ln(\varepsilon_\star), \end{array}$$

whence $\ln\left(\int_\Omega \exp\{-\phi_*^a(\omega)\} p_x(\omega) P(d\omega)\right) = \ln\left(\int_\Omega \exp\{-\phi_*(\omega)\} p_x(\omega) P(d\omega)\right) + a \leq \ln(\varepsilon_\star) + a$, and (56.$a$) follows. Similarly, when $y \in Y$, we have

$$\begin{array}{rcl} 2\ln(\varepsilon_\star) & = & \Phi_* \geq \Phi(\phi_*, [x_*; y]) \\ & = & \ln\left(\int_\Omega \exp\{-\phi_*(\omega)\} p_{x_*}(\omega) P(d\omega)\right) + \ln\left(\int_\Omega \exp\{\phi_*(\omega)\} p_y(\omega) P(d\omega)\right) \\ & = & \ln(\varepsilon_\star) + \ln\left(\int_\Omega \exp\{\phi_*(\omega)\} p_y(\omega) P(d\omega)\right), \end{array}$$

so that $\ln\left(\int_\Omega \exp\{\phi_*^a(\omega)\} p_y(\omega) P(d\omega)\right) = \ln\left(\int_\Omega \exp\{\phi_*(\omega)\} p_y(\omega) P(d\omega)\right) - a \leq \ln(\varepsilon_\star) - a$, and (56.$b$) follows.

Now let $x \in X$, and let $\epsilon(x)$ be the probability for the test, the detector being ϕ_*^a, to reject H_X; this is at most the probability for $\phi_*^a(\omega)$ to be nonpositive when $\omega \sim p_x(\cdot)$, and therefore

$$\epsilon(x) \leq \int_\Omega \exp\{-\phi_*^a(\omega)\} p_x(\omega) P(d\omega),$$

so that $\epsilon(x) \leq \exp\{a\}\varepsilon_\star$ by (56.a). Thus, the probability for our test to reject the hypothesis H_X when it is true is $\leq \exp\{a\}\varepsilon_\star$. Relation (56.$b$) implies in the same fashion that the probability for our test to reject H_Y when this hypothesis is true is $\leq \exp\{-a\}\varepsilon_\star$.

4^0. Theorem 2.1.ii is readily given by the following

Lemma A.2 *Under the premise of Theorem 2.1, let $(\phi_*, [x_*; y_*])$ be a saddle point of Φ, and let $\epsilon \geq 0$ be such that there exists a (whatever) test for deciding between two simple hypotheses*

$$(A) : \omega \sim p(\cdot) := p_{x_*}(\cdot), \quad (B) : \omega \sim q(\cdot) := p_{y_*}(\cdot) \tag{57}$$

with the sum of error probabilities $\leq 2\epsilon$. Then

$$\varepsilon_\star \leq 2\sqrt{(1-\epsilon)\epsilon}. \tag{58}$$

Proof. Under the premise of the lemma, (A) and (B) can be decided with the sum of error probabilities $\leq 2\epsilon$, and therefore the test affinity of (A) and (B) is bounded by 2ϵ:

$$\int_\Omega \min[p(\omega), q(\omega)] P(d\omega) \leq 2\epsilon.$$

On the other hand, we have seen that the saddle point value of Φ is $2\ln(\varepsilon_\star)$; since $[x_*; y_*]$ is a component of a saddle point of Φ, it follows that $\min_{\phi \in \mathcal{F}} \Phi(\phi, [x_*; y_*]) = 2\ln(\varepsilon_\star)$. The left hand side in this equality, as we know from item 1^0, is $\Phi(\phi_{x_*,y_*}, [x_*; y_*])$, and we arrive at $2\ln(\varepsilon_\star) = \Phi(\frac{1}{2}\ln(p_{x_*}(\cdot)/p_{y_*}(\cdot)), [x_*; y_*]) = 2\ln\left(\int_\Omega \sqrt{p_{x_*}(\omega) p_{y_*}(\omega)} P(d\omega)\right)$, so that $\varepsilon_\star = \int_\Omega \sqrt{p_{x_*}(\omega) p_{y_*}(\omega)} P(d\omega) = \int_\Omega \sqrt{p(\omega) q(\omega)} P(d\omega)$. We now have (cf. [37, chapter 4])

$$\begin{array}{l} \varepsilon_\star = \int_\Omega \sqrt{p(\omega) q(\omega)} P(d\omega) = \int_\Omega \sqrt{\min[p(\omega), q(\omega)]} \sqrt{\max[p(\omega), q(\omega)]} P(d\omega) \\ \leq \left(\int_\Omega \min[p(\omega), q(\omega)] P(d\omega)\right)^{1/2} \left(\int_\Omega \max[p(\omega), q(\omega)] P(d\omega)\right)^{1/2} \leq \sqrt{2(2-2\epsilon)\epsilon} = 2\sqrt{(1-\epsilon)\epsilon}. \end{array}$$

5^0. We have proved items (i) and (ii) of Theorem 2.1. To complete the proof of the theorem, it remains to justify (6). Thus, let $(\phi_*, [x_*; y_*])$ be a saddle point of Φ satisfying (54). All we need to prove is that ϕ_* is nothing but

$$\bar{\phi}(\cdot) = \frac{1}{2}\ln\left(p_{x_*}(\cdot)/p_{y_*}(\cdot)\right).$$

Indeed, the function $\Phi(\cdot, [x_*; y_*])$ attains its minimum on $\mathcal{F}$ at the point ϕ_*; by Remark A.1, it follows that $\phi_*(\cdot) - \bar{\phi}(\cdot)$ is constant on Ω; since both $\bar{\phi}$ and ϕ_* satisfy (54), this constant is zero. □

A.2 Proofs of Propositions 3.1 and 3.2

Proposition 3.1 is a simple particular case of Proposition 3.2 which we prove here.

Observe that when $t \le K$ and $p \in X_t$, so that $p \in X_{it}$ for some $i \in \mathcal{I}_t$, we have by definition of ϕ_t, see (24),

$$\begin{aligned}
&\int_{\Omega_t} \exp\{-\phi_t(\omega_t)\}p(\omega_t)P_t(d\omega_t) = \int_{\Omega_t} \exp\{\min_{r\in\mathcal{I}_t}\max_{s\in\mathcal{J}_t}[a_{rst} - \phi_{rst}(\omega_t)]\}p(\omega_t)P_t(d\omega_t)\\
&\le \int_{\Omega_t} \exp\{\max_{s\in\mathcal{J}_t}[a_{ist} - \phi_{ist}(\omega_t)]\}p(\omega_t)P_t(d\omega_t) \le \sum_{s\in\mathcal{J}_t}\int_{\Omega_t} \exp\{a_{ist} - \phi_{ist}(\omega_t)\}p(\omega_t)P_t(d\omega_t)\\
&\le \sum_{s\in\mathcal{J}_t} \exp\{a_{ist}\}\epsilon_{ist} = \sum_{s\in\mathcal{J}_t} h_s^t\epsilon_{ist}/g_i^t \quad [\text{see (21.}a\text{), (24)}]\\
&= [E_t h^t]_i/g_i^t = \varepsilon_t \quad [\text{see (23)}].
\end{aligned} \tag{59}$$

Similarly, when $t \le K$ and $p \in Y_t$, so that $p \in Y_{jt}$ for some $j \in \mathcal{J}_t$, we have

$$\begin{aligned}
&\int_{\Omega_t} \exp\{\phi_t(\omega_t)\}p(\omega_t)P_t(d\omega_t) = \int_{\Omega_t} \exp\{\max_{r\in\mathcal{I}_t}\min_{s\in\mathcal{J}_t}[\phi_{rst}(\omega_t) - a_{rst}]\}p(\omega_t)P_t(d\omega_t)\\
&\le \int_{\Omega_t} \exp\{\max_{r\in\mathcal{I}_t}[\phi_{rjt}(\omega_t) - a_{rjt}]\}p(\omega_t)P_t(d\omega_t) \le \sum_{r\in\mathcal{I}_t}\int_{\Omega_t} \exp\{\phi_{rjt}(\omega_t) - a_{rjt}\}p(\omega_t)P_t(d\omega_t)\\
&\le \sum_{r\in\mathcal{I}_t} \exp\{-a_{rjt}\}\epsilon_{rjt} = \sum_{r\in\mathcal{I}_t} g_r^t\epsilon_{rjt}/h_j^t \ [\text{see (21.}b\text{), (24)}]\\
&= [E_t^T g^t]_j/h_j^t = \varepsilon_t \ [\text{see (23)}].
\end{aligned} \tag{60}$$

Now let $H_1 = H_X$ be true, let $\mathbf{E}_{|\zeta_1^{t-1}}\{\cdot\}$ stand for the conditional expectation, ζ_1^{t-1} being fixed, and let $p_{\zeta_1^{t-1}}(\cdot)$ be conditional, ζ_1^{t-1} being fixed, probability density of ω_t w.r.t. P_t, so that $p_{\zeta_1^{t-1}}(\cdot) \in X_t$ for all ζ_1^{t-1} and all $t \le K$. We have

$$\begin{aligned}
&\mathbf{E}\left\{\exp\{-\phi_1(\omega_1) - ... - \phi_t(\omega_t)\}\right\} = \mathbf{E}\left\{\exp\{-\phi_1(\omega_1) - ... - \phi_{t-1}(\omega_{t-1})\}\mathbf{E}_{|\zeta_1^{t-1}}\{\exp\{-\phi_t(\omega_t)\}\}\right\}\\
&= \mathbf{E}\left\{\exp\{-\phi_1(\omega_1) - ... - \phi_{t-1}(\omega_{t-1})\}\int_{\Omega_t}\exp\{-\phi_t(\omega_t)\}p_{\zeta_1^{t-1}}(\omega_t)P_t(d\omega_t)\}\right\}\\
&\le \varepsilon_t\mathbf{E}\left\{\exp\{-\phi_1(\omega_1) - ... - \phi_{t-1}(\omega_{t-1})\}\right\},
\end{aligned}$$

where the concluding inequality is due to (59). From the resulting recurrence,

$$\mathbf{E}\{\exp\{-\phi^K(\omega^K)\}\} \le \prod_{t=1}^K \varepsilon_t.$$

This inequality combines with the description of our test to imply that the probability to reject H_X when it is true is at most $\prod_{t=1}^K \varepsilon_t$.

Now assume that $H_2 = H_Y$ holds true, so that the conditional, ζ_2^{t-1} being fixed, distribution $p_{\zeta_2^{t-1}}(\cdot)$ of ω_t belongs to Y_t for all ζ_2^{t-1} and all $t \le K$. Applying the previous reasoning to $-\phi^K$ in the role of ϕ^K, ζ_2^t in the role of ζ_1^t, and (60) in the role of (59), we conclude that the probability to reject H_Y when it is true is at most $\prod_{t=1}^K \varepsilon_t$. □

A.3 Proof of Proposition 3.3

1^0. The matrix $\bar{E} = [p_i\epsilon_{ij}]_{1\le i,j\le m}$ has zero diagonal and positive off-diagonal entries. By the Perron-Frobenius theorem, the largest in magnitude eigenvalue of $\bar{E}$ is some positive real ρ, and the corresponding eigenvector g can be selected to be nonnegative. In addition, $g \ge 0$ is in fact positive, since the relation

$$\rho g_i = [\bar{E}g]_i$$

along with the fact the all p_i and all off-diagonal entries in E are positive, allows for $g_i = 0$ only if all the entries g_j with $j \ne i$ are zeros, that is, only when $g = 0$, which is impossible. Since $g > 0$, we can set

$$\alpha_{ij} = \bar{\alpha}_{ij} := \ln(g_j) - \ln(g_i),$$

thus ensuring $\alpha_{ij} = -\alpha_{ji}$ and

$$p_i\varepsilon_i = \sum_{j=1}^m p_i\epsilon_{ij}\exp\{\alpha_{ij}\} = \sum_{j=1}^m p_i\epsilon_{ij}g_j/g_i = g_i^{-1}\sum_{j=1}^m p_i\epsilon_{ij}g_j = g_i^{-1}[\bar{E}g]_i = \rho.$$

Thus, with our selection of α_{ij} we get

$$\varepsilon = \rho.$$

2^0. We claim that in fact $\varepsilon_* = \rho$, that is, the feasible solution $[\bar{\alpha}_{ij}]$ is optimal for (29). Indeed, otherwise there exists a feasible solution $[\alpha_{ij} = \bar{\alpha}_{ij} + \delta_{ij}]_{i,j}$ with $\delta_{ij} = -\delta_{ji}$ such that

$$\bar{\rho} = \max_i \left[p_i \sum_j \epsilon_{ij}\exp\{\alpha_{ij}\}\right] < \rho.$$

As we have shown, for every i we have $\rho = \sum_j p_i\epsilon_{ij}\exp\{\bar{\alpha}_{ij}\}$. It follows that the convex functions

$$f_i(t) = \sum_j p_i\epsilon_{ij}\exp\{\bar{\alpha}_{ij} + t\delta_{ij}\}$$

all are equal to ρ when $t = 0$ and are $\le \bar{\rho} < \rho$ when $t = 1$, whence, due to convexity of f_i, for every i one has

$$0 > \frac{d}{dt}\Big|_{t=0} f_i(t) = \sum_j p_i\epsilon_{ij}\exp\{\bar{\alpha}_{ij}\}\delta_{ij} = p_i\sum_j g_jg_i^{-1}\epsilon_{ij}\delta_{ij}.$$

Multiplying the resulting inequalities by $g_i^2/p_i > 0$ and summing up the results over i, we get

$$0 > \sum_{i,j} g_ig_j\epsilon_{ij}\delta_{ij}.$$

This is impossible, since $\epsilon_{ij} = \epsilon_{ji}$ and $\delta_{ij} = -\delta_{ji}$, and the right hand side in the latter inequality is zero. □

A.4 Proof of Proposition 3.4

In the notation and under the premise of the proposition, let $\widehat{\epsilon}_{ij}$ be the risks of detectors ϕ_{ij} as defined in Theorem 2.1, so that $\widehat{\epsilon}_{ij}^K$ are the risks of ϕ_{ij}^K. Denote δ the maximum of the risks $\widehat{\epsilon}_{ij}$ taken over all "far from each other" pairs of indexes (i,j), that is, pairs such that i,j do not belong to the same group $\mathcal{I}_\ell$, $\ell=1,...,L$, and let $\bar{i},\bar{j}$ be two "far from each other" indexes such that $\delta=\widehat{\epsilon}_{\bar{i}\bar{j}}$. Test $\overline{T}$ clearly induces a test for deciding on the pair of hypotheses $H^1:=H_{\bar{i}}$, $H^2:=H_{\bar{j}}$ from observation $\omega^{\bar{K}}$ which does not accept H^χ, $\chi=1,2$, when the hypothesis is true, with probability at most ϵ, and never accepts both these hypotheses simultaneously. Same as in the proof of Proposition 2.1, the latter implies that $\delta^{\bar{K}}=[\widehat{\epsilon}_{\bar{i}\bar{j}}]^{\bar{K}}\le 2\sqrt{\epsilon}$. Since the nonzero entries in the matrix $D=D_K$ participating in the description of the test $\widehat{\mathcal{T}}^K$ are of the form $\widehat{\epsilon}_{ij}^K$ with "far from each other" i,j, the entries in the entrywise nonnegative matrix D_K do not exceed $\delta^K\le[2\sqrt{\epsilon}]^{K/\bar{K}}$. Therefore the spectral norm of D_K (which, as we know, upper bounds the risk of $\widehat{\mathcal{T}}^K$) does not exceed $M[2\sqrt{\epsilon}]^{K/\bar{K}}$, and the conclusion of Proposition 3.4 follows. □

A.5 Proofs of Propositions 4.1 and 4.2

We prove here Proposition 4.1, the proof of Proposition 4.2 can be conducted following same lines.

1^0. Let us fix i. It is immediately seen that problem (P_ϵ^i) is solvable (recall that $Ae[i]\neq 0$); let $\rho^i=\rho_i^P(\epsilon)$, r^i, u^i, v^i be an optimal solution to this problem. We clearly have $r^i=\rho^i$. We claim that the optimal value in the optimization problem

$$\min_{r,u,v}\left\{\frac{1}{2}\sum_\ell\left[\sqrt{[Au]_\ell}-\sqrt{[A(re[i]+v^i)]_\ell}\right]^2 : u\in\mathcal{V}, v\in\mathcal{V}, \rho^i\le r\le R\right\} \qquad (P)$$

is $\ln(\sqrt{n}/\epsilon)$, while (r^i,u^i,v^i) is an optimal solution to the problem. Indeed, taking into account the origin of $u^i,v^i,\rho^i=r^i$ and the relation $R\ge\rho_i^P(\epsilon)$, (r^i,u^i,v^i) is a feasible solution to this problem with the value of the objective $\le\ln(\sqrt{n}/\epsilon)$; thus, all we need in order to support our claim is to verify that the optimal value in (P) is $\ge\ln(\sqrt{n}/\epsilon)$. To this end assume for a moment that (P) has a feasible solution $(\bar{r},\bar{u},\bar{v})$ with the value of the objective $<\ln(\sqrt{n}/\epsilon)$. Then, setting $\rho^+=\rho^i+\delta$, $r^+=\bar{r}+\delta$, $u^+=\bar{u}$, $v^+=\bar{v}$ and choosing $\delta>0$ small enough, we clearly get a feasible solution to (P_ϵ^i) with the value of the objective $>\rho^i=\rho_i^P(\epsilon)$, which is impossible. Our claim is justified.

2^0. Recalling the "Poisson case" discussion in section 2, item 1^0 implies that the simple test associated with the detector $\phi_i(\cdot)$ given by (36) decides between the hypotheses H_0 and $H^i(\rho_i^P(\epsilon))$ with probabilities of errors $\le\epsilon/\sqrt{n}$. Since $H^i(r)$ "shrinks" as r grows, we conclude that whenever $\rho_i\in[\rho_i^P(\epsilon),R]$, the same test decides between the hypotheses H_0 and $H^i(\rho_i)$ with probabilities of errors not exceeding $\epsilon/\sqrt{n}$. Now let $\rho=[\rho_1;...;\rho_n]$ satisfy the premise of Proposition 4.1, so that $\rho_i\ge\rho_i^P(\epsilon)$ for all i. Note that the problem of testing $H_0:\mu\in X$ against $H_1(\rho):\mu\in\bigcup_{i=1}^n Y(\rho_i)$, along with the tests $\phi_{1i}(\cdot)=\phi_i(\cdot)$, $i=1,...,n$ satisfy the premise of Proposition 3.1 with $\epsilon_{1i}=\epsilon/\sqrt{n}$, $\varepsilon=\sqrt{\sum_{i=1}^n\epsilon_{1i}^2}(=\epsilon)$, and $a_{1i}=-\frac{1}{2}\ln n$, $i=1,...,n$. As a result, by Proposition 3.1, the risk of the test $\phi^P(\cdot)$ does not exceed ϵ.

3^0. To justify the bound on rate optimality, let us set

$$\mathrm{Opt}_i(\rho) = \min_{r,u,v}\left\{\frac{1}{2}\sum_\ell\left[\sqrt{[Au]_\ell} - \sqrt{[A(re[i]+v^i)]_\ell}\right]^2 : u\in\mathcal{V}, v\in\mathcal{V}, \rho\le r\le R\right\} \qquad [\rho\ge 0]$$

The function $\mathrm{Opt}(\rho)$ by its origin is a nondecreasing convex function on the segment $0\le\rho\le R$, $\mathrm{Opt}_i(\rho)=+\infty$ when $\rho>R$, and $\mathrm{Opt}(0)=0$. It follows that

$$\forall(\rho\in[0,R],\theta\ge 1): \mathrm{Opt}_i(\theta\rho)\ge\theta\mathrm{Opt}_i(\rho) \tag{61}$$

Now assume that for some $\rho=[\rho_1;...;\rho_n]$ and $\epsilon\in(0,1/4)$ there exists a test which decides between H_0 and $H_1(\rho)$ with probability of error $\le\epsilon$. Taking into account the union structure of $H_1(\rho)$, for every fixed i this test decides with the same probabilities of errors between the hypotheses H_0 and $H^i(\rho_i)$. All we need in order to prove the bound on the rate of optimality of $\widehat{\phi}_P$ is to extract from the latter observation that $\rho_i^P(\epsilon)/\rho_i\le\kappa_n:=\kappa_n(\epsilon)$ for every i. Let us fix i and verify that $\rho_i^P(\epsilon)/\rho_i\le\kappa_n$. There is nothing to do when $\rho_i\ge\rho_i^P(\epsilon)$ (due to $\kappa_n\ge 1$); thus, assume that $\rho_i<\rho_i^P(\epsilon)$. Note that $\rho_i>0$ (since otherwise the hypotheses H_0 and $H^i(\rho_i)$ have a nonempty intersection and thus cannot be decided with probabilities of errors $<1/2$, while we are in the case of $\epsilon<1/4$). Applying Theorem 2.1 to the pair of hypotheses H_0, $H^i(\rho_i)$, it is straightforward to see that in this case item (ii) of Theorem states exactly that $\exp\{-\mathrm{Opt}_i(\rho_i)\}\le 2\sqrt{\epsilon}$, or, which is the same, $\mathrm{Opt}(\rho_i)\ge\delta:=\frac{1}{2}\ln(1/\epsilon)-\ln(2)$; δ is positive due to $\epsilon\in(0,1/4)$. Now let $\theta>\ln(\sqrt{n}/\epsilon)/\delta$, so that $\theta\ge 1$. By (61), we either have $\theta\rho_i>R$, whence $\theta\rho_i\ge\rho_i^P(\epsilon)$ due to $\rho_i^P(\epsilon)\le R$, or $\theta\rho_i\le R$ and $\mathrm{Opt}_i(\theta\rho_i)>\ln(\sqrt{n}/\epsilon)$. In the latter case, as we have seen in item $\mathbf{1}^0$ of the proof, it holds $\mathrm{Opt}_i(\rho_i^P(\epsilon))=\ln(\sqrt{n}/\epsilon)$, and thus $\rho_i^P(\epsilon)<\theta\rho_i$ since Opt_i is nondecreasing in $[0,R]$. Thus, in all cases $\theta\rho_i>\rho_i^P(\epsilon)$ whenever $\theta>\ln(\sqrt{n}/\epsilon)/\delta$. But the latter ratio is exactly κ_n, and we conclude that $\kappa_n\rho_i\ge\rho_i^P(\epsilon)$, as required. □

A.6 Proof of Proposition 4.3

1^0. Let the premise in Proposition 4.3 hold true, and let us set $\varrho=\rho[\epsilon]$. Observe, first, that $\mathrm{Opt}[\varrho]=\ln\epsilon$. Indeed, problem (44) clearly is solvable, and $\bar{x},\bar{y},r=\varrho$ is an optimal solution to this problem. $(\bar{x},\bar{y})$ is a feasible solution to $(F_{g,\alpha}[\varrho])$, whence the optimal value in the latter problem is at least $\ln\epsilon$. Now let us lead to a contradiction the assumption that $\mathrm{Opt}[\varrho]>\ln\epsilon$. Under this assumption, let $x_0\in H_0[\rho_{\max}]$, $y_0\in H_1[\rho_{\max}]$, and let $(\hat{x},\hat{y})$ be an optimal solution to $(F_{g,\alpha}[\varrho])$, so that

$$\sum_{\ell=1}^L K_\ell\ln\left(\sum_{i=1}^{n_\ell}\sqrt{[A^\ell x]_i[A^\ell y]_i}\right)>\ln\epsilon \tag{62}$$

when $x=\hat{x}$, $y=\hat{y}$. Now let $x_t=\hat{x}+t(x_0-\hat{x})$, $y_t=\hat{y}+t(y_0-\hat{y})$. Since (62) hods true for $x=\hat{x}$, $y=\hat{y}$, for small enough positive t we have

$$g^Tx_t\le\alpha-\varrho-t(\rho_{\max}-\varrho),\ g^Ty_t\ge\alpha+\varrho+t(\rho_{\max}-\varrho),\ \sum_{\ell=1}^L K_\ell\ln\left(\sum_{i=1}^{n_\ell}\sqrt{[A^\ell x_t]_i[A^\ell y_t]_i}\right)\ge\ln\epsilon.$$

which, due to $\rho_{\max}>\varrho$, contradicts the fact that ϱ is the optimal value in (44).

2^0. Let us prove (46). This relation is trivially true when $\varrho = 0$, thus assume that $\varrho > 0$. Since $\rho_{\max} \geq 0$, and $g^T x$ takes on X both values $\leq \alpha$ and values $\geq \alpha$, this implies, by convexity of $\mathcal{X}$, that $g^T x$ takes value α somewhere on X. Therefore, the hypotheses $H_0[0]$ and $H_1[0]$ intersect, whence $\mathrm{Opt}[0] = 0$. In addition to this, due to its origin, $\mathrm{Opt}[\rho]$ is a concave function of $\rho \in [0, \varrho]$. Thus, $\mathrm{Opt}[\theta\varrho] \geq \theta \mathrm{Opt}[\varrho] = \theta \ln \epsilon$ when $0 \leq \theta \leq 1$. Now, to prove (46) is exactly the same as to prove that when $0 \leq \rho < \vartheta^{-1}(\epsilon)\varrho$, no test for problem $(\mathcal{D}_{g,\alpha}[\rho])$ with risk $\leq \epsilon$ is possible. Assuming, on the contrary, that $0 \leq \rho < \vartheta(\epsilon)\varrho$ and $(\mathcal{D}_{g,\alpha}[\rho])$ admits a test with risk $\leq \epsilon$; same as in the proof of Theorem 2.1.ii, this implies that for every $x \in H_0[\rho]$ and $y \in H_1[\rho]$, the Hellinger affinity of the distributions of observations associated with x and y does not exceed $2\sqrt{\epsilon}$, whence $\mathrm{Opt}[\rho] \leq \ln(2\sqrt{\epsilon})$. On the other hand, as we have seen, $\mathrm{Opt}[\rho] \geq \frac{\rho}{\varrho} \ln \epsilon$, and we arrive at $\frac{\rho}{\varrho} \ln \epsilon \leq \ln(2\sqrt{\epsilon})$, whence $\vartheta^{-1}(\epsilon) > \rho/\varrho \geq \frac{\ln(2\sqrt{\epsilon})}{\ln \epsilon} = \vartheta^{-1}(\epsilon)$, which is impossible.

3^0. Let now $\rho \in [\varrho, \rho_{\max}]$, so that problem $(F_{g,\alpha}[\rho])$ is solvable with optimal value $\mathrm{Opt}[\rho]$; clearly, $\mathrm{Opt}[\rho]$ is a nonincreasing function of ρ, whence $\mathrm{Opt}[\rho] \leq \mathrm{Opt}[\varrho] = \epsilon$, Applying Proposition 2.2 (with no Gaussian and Poisson factors and $a = 0$) and recalling the origin of $\mathrm{Opt}[\rho]$, we conclude that the risk of the simple test with the detector $\widehat{\phi}_\rho$ does not exceed $\exp\{\mathrm{Opt}[\rho]\} \leq \epsilon$. □