Hypothesis Testing via Euclidean Separation

Vincent Guigues * Anatoli Juditsky † Arkadi Nemirovski ‡

Abstract

We discuss an "operational" approach to testing convex composite hypotheses when the underlying distributions are heavy-tailed. It relies upon Euclidean separation of convex sets and can be seen as an extension of the approach to testing by convex optimization developed in [9, 13]. In particular, we show how one can construct quasi-optimal testing procedures for families of distributions which are majorated, in a certain precise sense, by a sub-spherical symmetric one and study the relationship between tests based on Euclidean separation and "potential-based tests." We apply the promoted methodology in the problem of sequential detection and illustrate its practical implementation in an application to sequential detection of changes in the input of a dynamic system.

Résumé

Nous proposons une méthode "opérationnelle" pour le problème de tests d'hypothèses composites convexes lorsque les distributions sous-jacentes possèdent des queues lourdes. Elle s'appuie sur la séparation euclidienne des ensembles convexes et peut être vue comme une extension de la méthode développée dans [9, 13] pour l'étude des tests d'hypothèses reposant sur des techniques d'optimisation convexe. En particulier, nous montrons comment construire des tests quasi-optimaux pour des familles de distributions qui sont majorées, dans un sens précis, par une distribution symétrique quasi-sphérique et étudions la relation entre les tests basés sur la séparation euclidienne et les tests utilisant des potentiels. Nous appliquons la méthodologie proposée au problème de la détection séquentielle et décrivons sa mise en oeuvre pour la détection séquentielle de ruptures dans l'entrée d'un système dynamique.

Keywords: Hypothesis testing, nonparametric testing, composite hypothesis testing, statistical applications of convex optimization.

1 Introduction

The following important observation, attributed to H. Chernoff [6] (see also [3, 4]), was the starting point of our research.

> Let X_1 and X_2 be two nonempty closed and convex sets, one of them being bounded, in $\mathbf{R}^n$. Suppose that, given a noisy observation
>
> $$\omega = x + \xi \tag{1}$$
>
> of the unknown signal $x \in X_1 \cup X_2$, where $\xi \sim \mathcal{N}(0, I_n)$ – the standard n-dimensional Gaussian vector, one wants to decide on the hypotheses $H_1 : x \in X_1$ vs. $H_2 : x \in X_2$. Then, assuming that X_1 and X_2 do not intersect (the decision problem is clearly unsolvable otherwise), optimal tests (with respect to different definitions of maximal risks) can be obtained using the following simple

*School of Applied Mathematics FGV/EMAp, 22 250-900 Rio de Janeiro, Brazil vincent.guigues@fgv.br

†LJK, Université Grenoble Alpes, 700 Avenue Centrale 38041 Domaine Universitaire de Saint-Martin-d'Hères, France, anatoli.juditsky@imag.fr

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

Research of the first author was partially supported by an FGV grant, CNPq grants 307287/2013-0, and 401371/2014-0, FAPERJ grant E-26/201.599/2014. Research of the second author was partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025) and CNPq grant 401371/2014-0. Research of the third author was partially supported by NSF grants CCF-1523768, CCF-1415498, and CNPq grant 401371/2014-0.

construction:
1) solve the following (convex optimization) Euclidean separation problem

$$\mathrm{Opt} = \min_{x^1 \in X_1,\, x^2 \in X_2} \tfrac{1}{2}\|x^1 - x^2\|_2 \tag{2}$$

where $\|\cdot\|_2$ is the Euclidean distance.
2) Given an optimal solution (x_*^1, x_*^2) to (2), compute

$$h_* = \frac{x_*^1 - x_*^2}{\|x_*^1 - x_*^2\|_2} \text{ and } c_* = \tfrac{1}{2} h_*^T (x_*^1 + x_*^2).$$

Then the test $\mathcal{T}_*$ which accepts H_1 if $h_*^T \omega - c_* \geq 0$ and accepts H_2 otherwise minimizes the maximal risk of testing – the maximal over $x \in X_1 \cup X_2$ probability of rejecting the true hypothesis – over the class of all (deterministic or randomized) tests. [1] Furthermore, the risk of $\mathcal{T}_*$ is easily computable: the maximal probability of wrongly rejecting the true hypothesis is

$$\frac{1}{\sqrt{2\pi}} \int_{\mathrm{Opt}}^{\infty} \mathrm{e}^{-t^2/2} dt = \mathrm{Erf}(\mathrm{Opt}),$$

where $\mathrm{Erf}(\cdot)$ is the standard error function.

This simple observation had important theoretical consequences (see [11, 15]). Surprisingly, its "practical implications" have been largely overseen. Indeed, when the problem (2) can be solved efficiently, [2] one can assemble pair-wise tests into multiple-testing procedures to build provably (nearly) optimal tests for a wide class of Gaussian decision problems (e.g., various detection problems [3, 4, 12, 7, 5]). Then in [9] the corresponding framework was extended to "good," in a certain precise sense, parametric families of distributions, which include, aside from the Gaussian family, families of Poisson and discrete distributions, the "common denominator" of these developments being the fact that for these families the *near-optimal* (plain optimal, in the case of Gaussian family) tests can be built upon using *affine detectors* which can be found by convex optimization. Later, *affine and quadratic detectors* were studied in a more general setting in [13].

In this paper we extend [9], [13] studying the application of tests based on "Euclidean separation" to the problems where the distribution of the observation noise ξ has "heavy tails." In particular, in Sections 2.2 and 2.3 we discuss the problem of testing convex hypotheses for *sub-spherical families* of distributions, which include Gaussian and Gaussian mixture distributions such as multivariate Student [14] and multivariate Laplace [8] distributions. We study the relationship of sub-spherical families with detector-based tests, and show how "good" detectors can be built for sub-spherical families, as well as for some other (e.g., sub-Gaussian) families of distributions in Section 2.4. Then in Section 3 we explain how tests based on Euclidean separation of pairs of convex hypotheses can be used to construct sequential change detection procedures. Finally, in Section 4 we present a numerical illustration of the proposed techniques: we implement sequential decision rules, developed in Sections 3.3 and 3.4 for a toy problem of detecting changes in the input of a dynamical system – changes in the trend of a simple time series.

2 Basic theory

2.1 Pairwise hypothesis testing: Situation and goal

The basic problem we intend to consider *in a nutshell* is as follows: we are given observation

$$\omega = x + \xi, \tag{3}$$

where $x \in \mathbf{R}^n$ is an unknown signal, and ξ is a random noise with probability density $p(\cdot)$, taken with respect to the Lebesgue measure, known to belong to some given family $\mathcal{P}$. Our basic goal is, given two nonempty

[1] A test which accepts H_1 if $h_*^T \omega - c_* \geq \alpha$ and accepts H_2 otherwise, with a properly chosen α, is also minimax optimal in several other settings, e.g., Neyman-Pearson problem, etc.

[2] what is the case when sets X_1 and X_2 allow for a "computationally efficient description." We refer the reader to [2] for precise definitions and details on efficient implementability. For the time being, it is sufficient to assume that (2) can be solved using `CVX` [10].

closed convex sets $X_\chi \subset \mathbf{R}^n$, $\chi = 1, 2$, with one of these sets bounded, to decide, via observation (3), on the hypotheses H_χ, $\chi = 1, 2$, with H_χ stating that the signal x underlying observation belongs to X_χ. In other words, we are to decide upon the families $\mathcal{P}_1$ and $\mathcal{P}_2$ of distributions, where $\mathcal{P}_\chi$ is the family of distributions of random vectors $x + \xi$, $\xi \sim p$ with $p \in \mathcal{P}$ and $x \in X_\chi$. For the sake of brevity, we shall simplify the description of the hypotheses H_χ to "$x \in X_\chi$."

Stationary repeated observations. We "embed" the just defined inference problem in the family of inference problems $(\mathcal{S}_K)$, $K = 1, 2, ...$, where $(\mathcal{S}_K)$ is the problem of deciding whether $x \in X_1$ or $x \in X_2$ via sample $\omega^K = (\omega_1, ..., \omega_K)$ of K independent observations

$$\omega_k = x + \xi_k,\ k = 1, ..., K, \tag{4}$$

with independent noises $\xi_k \sim p(\cdot) \in \mathcal{P}$; we refer to observations (4) with independent across k noises $\xi_k \sim p \in \mathcal{P}$ as to *stationary K-repeated observations*.

Semi-stationary repeated observations. Inference problem $(\mathcal{S}_K)$ can be viewed as a special case of a more general inference problem $(\overline{\mathcal{S}}_K)$ as follows. Suppose that $\{X_1^k, X_2^k : 1 \le k \le K\}$ is a collection of nonempty convex and closed sets such that at least one of the set in every pair (X_1^k, X_2^k) is bounded. Let also $\mathcal{P}^1, ..., \mathcal{P}^K$ be K given families of probability densities with respect to the Lebesgue measure. Suppose that the observation $\omega^K = (\omega_1, ..., \omega_K)$ is given by

$$\omega_k = x_k + \xi_k,\ k = 1, ..., K, \tag{5}$$

where $\{x_k\}_{k=1}^K$ is a deterministic sequence, $\xi_k \sim p_k$ are independent across k noises, and $\{p_k \in \mathcal{P}^k\}_{k=1}^K$ is a deterministic sequence.[3] In the sequel, we refer to observations (5) satisfying the just imposed restrictions on $\{x_k\}_{k=1}^K$ and $\{\xi_k\}_{k=1}^K$, as to *semi-stationary K-repeated observations*. Our objective in problem $(\overline{\mathcal{S}}_K)$ is to decide, via the observations (5), on the hypotheses H_χ, $\chi = 1, 2$, with H_χ stating that $x_k \in X_\chi^k$ for all $k = 1, 2, ..., K$.

A simple test for $(\mathcal{S}_K)$ or $(\overline{\mathcal{S}}_K)$ is, by definition, a function $\mathcal{T}_K(\omega^K)$, $\omega^K = (\omega_1, ..., \omega_K)$, taking values $\{1, 2\}$, with $\mathcal{T}_K(\omega^K) = \chi$ interpreted as "given observation ω^K, the test accepts H_χ and rejects the alternative." We define the *partial risks* $\mathrm{Risk}_{\chi\mathcal{S}}(\mathcal{T}_K|\mathcal{P}, X_1, X_2)$, $\chi = 1, 2$ of a test $\mathcal{T}_K$ on the inference problem $(\mathcal{S}_K)$ as

$$\begin{array}{rcl} \mathrm{Risk}_{1\mathcal{S}}(\mathcal{T}_K|\mathcal{P}, X_1, X_2) & = & \sup_{x\in X_1} \sup_{p(\cdot)\in\mathcal{P}} \mathrm{Prob}_{x,p}\{\mathcal{T}_K(\omega^K) = 2\}, \\ \mathrm{Risk}_{2\mathcal{S}}(\mathcal{T}_K|\mathcal{P}, X_1, X_2) & = & \sup_{x\in X_2} \sup_{p(\cdot)\in\mathcal{P}} \mathrm{Prob}_{x,p}\{\mathcal{T}_K(\omega^K) = 1\}, \end{array}$$

where $\mathrm{Prob}_{x,p}$ stands for the probability with respect to the distribution of observations (4). In other words, $\mathrm{Risk}_{\chi\mathcal{S}}(\mathcal{T}_K|\mathcal{P}, X_1, X_2)$, is the worst-case probability for $\mathcal{T}_K$ to reject H_χ when the hypothesis is true. The partial risks $\mathrm{Risk}_{\chi\overline{\mathcal{S}}}(\mathcal{T}_K|[\mathcal{P}^k, X_1^k, X_2^k]_{k=1}^K)$, $\chi = 1, 2$, of a test $\mathcal{T}_K$ on the inference problem $(\overline{\mathcal{S}}_K)$ are defined similarly, but now the supremum is taken with respect to *all* deterministic sequences $\{x_k \in X_\chi^k\}_{k=1}^K$ and $\{p_k \in \mathcal{P}^k\}_{k=1}^K$ participating in (5). Finally, the risks of $\mathcal{T}_K$ on $(\mathcal{S}_K)$ and $(\overline{\mathcal{S}}_K)$ are defined as

$$\begin{array}{rcl} \mathrm{Risk}_{\mathcal{S}}(\mathcal{T}_K|\mathcal{P}, X_1, X_2) & = & \max\limits_{\chi=1,2} \mathrm{Risk}_{\chi\mathcal{S}}(\mathcal{T}_K|\mathcal{P}, X_1, X_2), \\ \mathrm{Risk}_{\overline{\mathcal{S}}}\left(\mathcal{T}_K|[\mathcal{P}^k, X_1^k, X_2^k]_{k=1}^K\right) & = & \max\limits_{\chi=1,2} \mathrm{Risk}_{\chi\overline{\mathcal{S}}}\left(\mathcal{T}_K|[\mathcal{P}^k, X_1^k, X_2^k]_{k=1}^K\right). \end{array}$$

Our intention is to investigate the performance of specific tests stemming from "Euclidean Separation" of X_1 and X_2 (or X_1^k and X_2^k) to be described in a while.

2.2 Sub-spherical families of distributions

We shall be primarily interested in *sub-spherical families* $\mathcal{P}$.

[3] One can easily verify (cf. [9, Section 3.1.2]) that the constructions and results to follow remain intact when the assumptions on observations (5) are weakened to the assumption that x_k and ξ_k are random, and the conditional distribution of ξ_k, given $x_1, ..., x_k$ and $\xi_1, .., \xi_{k-1}$, always belongs to $\mathcal{P}^k$.

2.2.1 Sub-spherical families of distributions: definition and basic examples

Definition 2.1 A. *A sub-spherical family of distributions $\mathcal{P} = \mathcal{P}_\gamma^n$ on $\mathbf{R}^n$ is specified by an even probability density $\gamma(\cdot)$ on the axis such that γ is positive in a neighbourhood of the origin. $\mathcal{P}_\gamma^n$ is comprised of all probability densities $p(\cdot)$ on $\mathbf{R}^n$ such that $p(\cdot)$ is even, and*

$$\forall(e \in \mathbf{R}^n, \|e\|_2 = 1, \delta \geq 0): \int\limits_{e^T\xi \geq \delta} p(\xi)d\xi \leq P_\gamma(\delta) := \int\limits_{\delta}^{\infty} \gamma(s)ds. \tag{6}$$

that is, p-probability mass of a half-space not containing a neighbourhood of the origin is upper-bounded by $P_\gamma(\delta)$, where δ is the distance from the origin to the half-space, and

$$P_\gamma(r) = \int\limits_{r}^{\infty} \gamma(s)ds: \ \mathbf{R} \to [0,1]. \tag{7}$$

B. *A sub-spherical family $\mathcal{P} = \mathcal{P}_\gamma^n$ is called monotone, if it contains a cap – a spherically symmetric density $q(\xi) = f(\|\xi\|_2)$ where f is nonincreasing on the nonnegative axis and such that the induced by q density of the distribution of $e^T\xi$, $\|e\|_2 = 1$, is exactly $\gamma(\cdot)$. Note that whenever this is the case, $\gamma(\cdot)$ is nonincreasing on the nonnegative ray.*

C. *We call a function γ on the real axis nice, if γ is an even probability density which is continuous and is nonincreasing on the nonnegative ray. A sub-family $\mathcal{P}^n$ of a sub-spherical family $\mathcal{P}_\gamma^n$ is called completely monotone, if γ is nice, and for every $p(\cdot) \in \mathcal{P}^n$ and every $e \in \mathbf{R}^n$, $\|e\|_2 = 1$, the random scalar variable $e^T\xi$, $\xi \sim p$, has probability density $\gamma_{e,p}(\cdot)$, and this density is nice. Note that due to $\mathcal{P}^n \subset \mathcal{P}_\gamma^n$, it holds*

$$\int\limits_{\delta}^{\infty} \gamma_{e,p}(s)ds \leq \int\limits_{\delta}^{\infty} \gamma(s)ds \quad \forall \delta \geq 0. \tag{8}$$

In the sequel, we simplify the notation $\mathcal{P}_\gamma^n$ to $\mathcal{P}_\gamma$ when the value of n is clear from the context.

Example: Gaussian scale mixtures. Consider the situation where

$$\xi \sim \sqrt{Z}\eta, \tag{9}$$

where Z is a scalar a.s. positive random variable with given probability distribution $P_Z(t) = \mathrm{Prob}\{Z \leq t\}, t \geq 0$ such that $P_Z(0) = 0$, and $\eta \sim \mathcal{N}(0, \Theta)$ is a zero mean Gaussian n-dimensional random vector independent of Z with unknown a priori positive definite covariance matrix Θ which is known to be $\preceq I_n$. We refer to Θ as to *matrix parameter* of the distribution of ξ. Given a unit vector e and $\delta \geq 0$, we have

$$\begin{aligned}
\mathrm{Prob}\{e^T\xi \geq \delta\} &= \int\limits_{t>0} \left[\mathrm{Prob}\{e^T\eta \geq t^{-1/2}\delta\}\right] dP_Z(t) \\
&= \int\limits_{t>0} \left[\mathrm{Prob}_{\zeta\sim\mathcal{N}(0,I_n)}\{e^T\Theta^{1/2}\zeta \geq t^{-1/2}\delta\}\right] dP_Z(t) \\
&= \int\limits_{t>0} \mathrm{Erf}\left(t^{-1/2}\delta/\sqrt{e^T\Theta e}\right) dP_Z(t) \leq \int\limits_{t>0} \mathrm{Erf}(t^{-1/2}\delta) dP_Z(t),
\end{aligned}$$

(here $\mathrm{Erf}(\cdot)$ is the standard error function), implying that

$$\mathrm{Prob}\{e^T\xi \geq \delta\} = \int\limits_{\delta}^{\infty} \gamma_{e,\Theta,P_Z}(s)ds \leq \int\limits_{\delta}^{\infty} \gamma_{P_Z}(s)ds \ \ \forall \delta \geq 0,$$

where

$$\begin{aligned}
\gamma_{e,\Theta,P_Z}(s) &= \int\limits_{t>0} \frac{1}{\sqrt{2\pi t e^T\Theta e}} \exp\left\{-\frac{s^2}{2te^T\Theta e}\right\} dP_Z(t), \\
\gamma_{P_Z}(s) &= \int_{t>0} \frac{1}{\sqrt{2\pi t}} \exp\left\{-\frac{s^2}{2t}\right\} dP_Z(t), \ -\infty < s < \infty.
\end{aligned}$$

Clearly, $\gamma_{e,\Theta,P_Z}(s)$ and $\gamma_{P_Z}(s)$ are positive even probability densities nonincreasing on the nonnegative ray; these densities are continuous, provided

$$\int_{t>0} t^{-1/2} dP_Z(t) < \infty. \tag{10}$$

Thus, the family of distributions of random vectors (9) with Z and η as explained above (we refer to these distributions as to *Gaussian scale mixtures*) is contained in the sub-spherical family $\mathcal{P}_{\gamma_{P_Z}}$. The latter family clearly is monotone, the cap being the probability density of $\sqrt{Z}\eta$ with independent $Z \sim P_Z$ and $\eta \sim \mathcal{N}(0, I_n)$. Besides this, in the case of (10) the family $\mathcal{P}_{\gamma_{P_Z}}$ is completely monotone.

A standard example of a Gaussian scale mixture is given by *n-variate t-distributions* $t_n(q, \Theta)$, $\Theta \preceq I_n$ (multivariate Student distributions with q degrees of freedom, see [14] and references therein). Here $t_n(q, \Theta)$ is, by definition, the distribution of the random vector $\xi = \sqrt{Z}\eta$ with $Z = q/\zeta$, where ζ is the independent of $\eta \sim \mathcal{N}(0, \Theta)$ random variable following χ^2-distribution with q degrees of freedom. One can easily see that all one-dimensional projections $e^T\xi$, $\|e\|_2 = 1$, of $\xi \sim t_n(q, I_n)$ are random variables with univariate t_q-distribution, implying that the multidimensional densities in question form a completely monotone sub-family of the sub-spherical family $\mathcal{P}_{\gamma_S}$ where γ_S is the density of Student's t_q distribution with q degrees of freedom.

Another example of scheme (9) is the *n-variate Laplace distributions* $\mathcal{L}_n(\lambda, \Theta)$, $\Theta \preceq I_n$, where Z is exponentially distributed with parameter λ. In this case all one-dimensional projections $e^T\xi$, $\|e\|_2 = 1$, of $\xi \sim \mathcal{L}_n(\lambda, I_n)$, obey the Laplace distribution with parameter λ, whence the distributions in question form a completely monotone sub-family of the sub-spherical family $\mathcal{P}_{\gamma_\mathcal{L}}$, where $\gamma_\mathcal{L}$ is the Laplace density

$$\gamma_\mathcal{L}(s) = (2\lambda)^{-1}\mathrm{e}^{-|s|/\lambda},\ s \in \mathbf{R}. \tag{11}$$

Finally, with Z taking value 1 with probability 1, scheme (9) describes Gaussian distributions with zero mean and covariance matrices $\preceq I_n$; all these distributions form a completely monotone sub-famuily of the sub-spherical family $\mathcal{P}_{\gamma_\mathcal{G}}$, where $\gamma_\mathcal{G}$ is the standard univariate Gaussian density:

$$\gamma_\mathcal{G}(s) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-s^2/2}.$$

2.2.2 "Calculus" of sub-spherical families of distributions

Sub-spherical families of distributions and their completely monotone subfamilies admit a kind of "calculus" with the basic rules which follow.

The following two facts are immediate:

Proposition 2.1 *Let $\mathcal{P}^n_\gamma$ be a sub-spherical family of distributions, and let $x \mapsto Qx : \mathbf{R}^n \to \mathbf{R}^m$ be an onto mapping satisfying $QQ^T \preceq I_m$. Whenever $p(\cdot) \in \mathcal{P}^n_\gamma$, the distribution of the random vector $Q\xi$, $\xi \sim p$, belongs to $\mathcal{P}^m_\gamma$. Moreover, if $QQ^T = I_m$ and $\mathcal{P}^n_\gamma$ has a cap q, then $\mathcal{P}^m_\gamma$ has a cap as well; this cap is the density of the random vector $Q\xi$, $\xi \sim q$.*

Proposition 2.2 *A sub-spherical family of distributions is closed with respect to taking convex combinations of its members. Besides this, the union $\mathcal{P}_{\gamma_1} \cup \mathcal{P}_{\gamma_2}$ of two sub-spherical families of distributions is contained in the sub-spherical family $\mathcal{P}_\gamma$ with*

$$\gamma(-s) = \gamma(s) = -\frac{d}{ds}\max_{i=1,2}\int_s^\infty \gamma_i(r)dr,\ s \geq 0.$$

Complete monotonicity is preserved by taking sums. The precise statement is as follows:

Proposition 2.3 *Let μ and ν be nice, let $\mathcal{P}^n$ be a subfamily of the sub-spherical family $\mathcal{P}^n_\mu$, and let $\mathcal{P}^m$ be a completely monotone subfamily of the sub-spherical family $\mathcal{P}^m_\nu$. Given $r \times n$ matrix A, $r \times m$ matrix B and positive definite matrix Θ such that*

$$\Theta^2 \succeq AA^T,\ \Theta^2 \succeq BB^T,\ AA^T + BB^T \succ 0, \tag{12}$$

consider random vectors of the form

$$\xi = \Theta^{-1}[A\eta + B\zeta], \tag{13}$$

where $\eta \sim p(\cdot) \in \mathcal{P}^n$ *and* $\zeta \sim q(\cdot) \in \mathcal{P}^m$ *are independent. Let also*

$$\gamma(s) = \int_{-\infty}^{\infty} \mu(s-r)\nu(r)dr. \tag{14}$$

Then

(i) $\gamma(\cdot)$ *is nice, and for every* $e \in \mathbf{R}^r$, $\|e\|_2 = 1$, *and every* $\delta \geq 0$ *one has*

$$\mathrm{Prob}\{e^T\xi \geq \delta\} \leq \int_{\delta}^{\infty} \gamma(s)ds. \tag{15}$$

Besides this, the scalar random variable $e^T\xi$ *possesses symmetric density, which combines with* (15) *to imply that the distribution of* ξ *belongs to the sub-spherical family* $\mathcal{P}^r_\gamma$.

(ii) If, in addition to the above assumptions, $\mathcal{P}^n$ *is completely monotone, then the family of distributions of random variables* (13) *induced by* $\eta \sim p \in \mathcal{P}^n$ *and* $\zeta \sim q \in \mathcal{P}^m$ *is a completely monotone subfamily of the sub-spherical family* $\mathcal{P}^r_\gamma$.

For a proof, see Section A.1.

As an immediate consequence of Proposition 2.3, we get the following

Corollary 2.1 *For* $1 \leq i \leq I < \infty$, *let* μ_i *be nice functions on the axis, and let* $\mathcal{P}^{n_i}$ *be subfamilies of the sub-spherical families* $\mathcal{P}^{n_i}_{\mu_i}$ *such that at least* $I-1$ *of these subfamilies are completely monotone. Given* $r \times n_i$ *matrices* A_i *such that* $A_iA_i^T + A_jA_j^T \succ 0$ *whenever* $i \neq j$, *let* $\Theta \succ 0$ *be such that*

$$\Theta^2 \succeq A_iA_i^T,\ 1 \leq i \leq I.$$

Let $\mathcal{P}^r$ *be the family of probability distributions of random vectors of the form*

$$\xi = \Theta^{-1}\sum_{i=1}^{I} A_i\eta_i$$

where $\eta_1, ..., \eta_I$ *are independent of each other and such that* $\eta_i \sim p_i$ *for some* $p_i \in \mathcal{P}^{n_i}$. *Then* $\mathcal{P}^r$ *is contained in the sub-spherical family* $\mathcal{P}^r_\gamma$, *where*

$$\gamma = \mu_1 \star \mu_2 \star ... \star \mu_I$$

is nice. If all $\mathcal{P}^{n_i}$ *are completely monotone, then so is* $\mathcal{P}^r$.

2.3 Euclidean separation and associated tests

In the sequel, we fix the entities $\mathcal{P}$, X_1, X_2, H_1, H_2 introduced in the beginning of Section 2.1 and assume that $X_1 \cap X_2 = \emptyset$ (otherwise no test can decide on H_1 vs. H_2 with risk $< 1/2$).

2.3.1 Single-observation Euclidean separation test

Consider the optimization problem

$$\mathrm{Opt} = \min_{x^1 \in X_1, x^2 \in X_2} \|x^1 - x^2\|_2, \tag{16}$$

and let x^1_*, x^2_* form an optimal solution to the problem (since both X_1, X_2 are nonempty, closed and convex, and one of the sets is bounded, an optimal solution does exist). We set

$$s_*(\omega) = h_*^T\omega - c_*,\ h_* = \frac{[x^1_* - x^2_*]}{\|x^1_* - x^2_*\|_2},\ c_* = \tfrac{1}{2}\left[\min_{x\in X_1} h_*^Tx + \max_{x\in X_2} h_*^Tx\right] = \tfrac{1}{2}h_*^T(x^1_* + x^2_*). \tag{17}$$

Note that while (16) may have many optimal solutions, the vector h_* and the real c_* are uniquely defined by X_χ, $\chi = 1, 2$. The affine function $s_*(\cdot)$ possesses the following properties:

$$\begin{array}{rcl} x \in X_1 & \Rightarrow & s_*(x) \geq s_*(x_*^1) = \delta, \ \ \delta := \frac{1}{2}\|x_*^1 - x_*^2\|_2, \\ x \in X_2 & \Rightarrow & s_*(x) \leq s_*(x_*^2) = -\delta. \end{array} \tag{18}$$

We can associate with h_* and c_* the *Euclidean separation test* $\mathcal{T}_1$ which, given observation ω (3) with x known to belong to $X_1 \cup X_2$, accepts the hypothesis $H_1 : x \in X_1$ and rejects the hypothesis $H_2 : x \in X_2$ when $s_*(\omega) \geq 0$, and accepts H_2 and rejects H_1 otherwise.

Let us make the following immediate observation:

Proposition 2.4 *In the situation of Section 2.1 and in the notation from (17), (18), let $\mathcal{P} = \mathcal{P}_\gamma$ be a sub-spherical family of distributions, and let $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$ be such that $\alpha_1 + \alpha_2 \leq 2\delta$. Whenever $x \in X_1$ and $p \in \mathcal{P}$, the p-probability of the event $\{\xi : s_*(x+\xi) < \frac{1}{2}(\alpha_2 - \alpha_1)\}$ is at most $P_\gamma(\alpha_1)$, and when $x \in X_2$, the p-probability of the event $\{\xi : s_*(x+\xi) \geq \frac{1}{2}(\alpha_2 - \alpha_1)\}$ is at most $P_\gamma(\alpha_2)$. As a result, risks of the test $\mathcal{T}_1$ satisfy*

$$\begin{array}{l} \mathrm{Risk}_{1\mathcal{S}}(\mathcal{T}_1|\mathcal{P}, X_1, X_2) \leq P_\gamma(\alpha_1), \\ \mathrm{Risk}_{2\mathcal{S}}(\mathcal{T}_1|\mathcal{P}, X_1, X_2) \leq P_\gamma(\alpha_2). \end{array}$$

In particular, when $\alpha_1 = \alpha_2 = \delta$, the risk $\mathrm{Risk}_{\mathcal{S}}(\mathcal{T}_1|\mathcal{P}, X_1, X_2)$ of the test is at most

$$\varepsilon_\star = \varepsilon_\star(\delta|\gamma) := P_\gamma(\delta) = \int_\delta^\infty \gamma(s)ds. \tag{19}$$

For a proof, see Section A.2.

Remark 2.1 *In the situation of Proposition 2.4, let the sub-spherical family $\mathcal{P}_\gamma$ be monotone. Then the test $\mathcal{T}_1$ described in the proposition has the minimal risk among all single-observation tests deciding on H_1 vs. H_2.*

Indeed, denoting by $q(\cdot)$ the cap of $\mathcal{P}_\gamma$, and by $p_1(\cdot)$ and $p_2(\cdot)$ the densities of random vectors $x_*^1 + \xi$, $x_*^2 + \xi$, $\xi \sim q(\cdot)$, we clearly have

$$\int \min[p_1(\omega), p_2(\omega)]d\omega = 2\varepsilon_\star,$$

implying by the Neyman-Pearson lemma that the risk of any test deciding on two simple hypotheses $x = x_1^*$, $x = x_2^*$ via a single observation (3) is at least $\varepsilon_\star$.

2.3.2 Majority tests based on Euclidean separation

Let K be a positive integer, and let $\mathcal{P} = \mathcal{P}_\gamma$ be a sub-spherical family of distributions. The *K-observation majority test* $\mathcal{T}_K^{\mathrm{maj}}$ for problem $(\mathcal{S}_K)$ works as follows: given observations ω_k, $k = 1, ..., K$, see (4), the test accepts H_1 and rejects H_2 when $s_*(\omega_k) \geq 0$ for at least $K/2$ values of k, and accepts H_2 and rejects H_1 otherwise. From Proposition 2.4 it follows that the risk of $\mathcal{T}_K^{\mathrm{maj}}$ satisfies the bound

$$\mathrm{Risk}_{\mathcal{S}}(\mathcal{T}_K^{\mathrm{maj}}|\mathcal{P}, X_1, X_2) \leq \sum_{K \geq k \geq K/2} \binom{K}{k} {\varepsilon_\star}^k (1 - \varepsilon_\star)^{K-k}. \tag{20}$$

Let us assume that in the problem $(\overline{\mathcal{S}}_K)$ the families $\mathcal{P}^1, ..., \mathcal{P}^K$ in the definition of $(\overline{\mathcal{S}}_K)$ are sub-spherical families of distributions, and that the sets X_1^k and X_2^k do not intersect for $k = 1, ..., K$. The majority test $\mathcal{T}_K^{\mathrm{maj}}$ can be easily modified to become applicable to problem $(\overline{\mathcal{S}}_K)$. Let the affine function $s_k(\cdot)$ and positive real δ_k be the entities $s_*(\cdot), \delta$ associated, via (16) – (18), with $\mathcal{P}^k$ in the role of $\mathcal{P}$ and X_χ^k in the role of X_χ, $\chi = 1, 2$. The *K-observation majority test* $\mathcal{T}_K^{\mathrm{maj}}$ for the problem $(\overline{\mathcal{S}}_K)$ given observations ω_k, $k = 1, ..., K$, see (5), accepts H_1 and rejects H_2 when $s_k(\omega_k) \geq 0$ for at least $K/2$ values of k, and accepts H_2 and rejects H_1 otherwise. Let now γ_k be the density underlying the sub-spherical family $\mathcal{P}^k$. When applying Proposition 2.4 to $\mathcal{P} = \mathcal{P}^k$, $X_1 = X_1^k$, and $X_2 = X_2^k$, we conclude that the risk $\mathrm{Risk}_{\mathcal{S}}(\mathcal{T}_1|\mathcal{P}^k, X_1^k, X_2^k)$ of the test $\mathcal{T}_1$ in the problem of deciding, *given a single observation* ω_k, upon the hypotheses $H_1^k : x_k \in X_1^k$ vs $H_2^k : x_k \in X_2^k$, does

not exceed $\epsilon_k(\delta_k) = \int_{\delta_k}^{\infty} \gamma_k(s)ds < \frac{1}{2}$. We conclude that the risk of the majority test $\mathcal{T}_K^{\rm maj}$ in the problem $(\overline{\mathcal{S}}_K)$ satisfies the bound [1]

$$\mathrm{Risk}_{\overline{\mathcal{S}}}\left(\mathcal{T}_K^{\rm maj}|[\mathcal{P}^k, X_1^k, X_2^k]_{k=1}^K\right) \le \sum_{K/2\le k\le K} p_{k|K}, \tag{21}$$

where $p_{k|K} = p_{k|K}(\epsilon_1(\delta_1), ..., \epsilon_K(\delta_K))$ is the probability of k successes in the first K non-stationary independent Bernoulli trials with the probability $\epsilon_k(\delta_k)$ of success in k-th trial. Observe that $p_{j|k}$, $0 \le j \le k$, $1 \le k \le K$, satisfy the recursion:

$$p_{j|k} = (1-\epsilon_k)p_{j|k-1} + \epsilon_k p_{j-1|k-1}, \text{ with } p_{0|0} = 1, \tag{22}$$

and where, by convention, $p_{-1|k} = 0$, $k = 0, ..., K-1$. Recursion (22) allows to compute all the quantities $p_{j|k}$, $0 \le j \le k \le K$, and thus the right hand side of (21), in $O(K^2)$ arithmetic operations.

2.4 Potential-based tests with Euclidean separation

2.4.1 Potentials and potential-based tests

Definition 2.2 *A potential is an odd and nondecreasing Borel real-valued function $\eta(\cdot)$ on the axis. Given a family $\mathcal{P}$ of probability densities on $\mathbf{R}^n$, a nonnegative δ and a potential $\eta(\cdot)$, we define the δ-risk $\mathrm{risk}_\delta(\eta|\mathcal{P})$ of the potential on $\mathcal{P}$ as the smallest ϵ such that*

$$\begin{array}{lrcl} (a) & \int \mathrm{e}^{-\eta(\delta+e^T\xi)}p(\xi)d\xi & \le & \epsilon \ \ \forall(e\in\mathbf{R}^n : \|e\|_2 = 1, p\in\mathcal{P}), \\ (b) & \int \mathrm{e}^{\eta(e^T\xi-\delta)}p(\xi)d\xi & \le & \epsilon \ \ \forall(e\in\mathbf{R}^n : \|e\|_2 = 1, p\in\mathcal{P}). \end{array} \tag{23}$$

Let us make the following immediate observation:

Proposition 2.5 *For $k = 1, ..., K$, let $\mathcal{P}^k$ be a family of probability densities on $\mathbf{R}^n$, X_1^k and X_2^k be closed nonempty non-intersecting convex sets in $\mathbf{R}^n$, one of the sets being bounded, and let h_k, c_k and δ_k be associated with $\mathcal{P} = \mathcal{P}^k$, $X_\chi = X_\chi^k$, $\chi = 1, 2$, via (17) and (18). Given potentials $\eta_1, ..., \eta_K$, let us define the Euclidean detector induced by the potential η_k and h_k, c_k as the function*

$$\phi_k(\omega) = \eta_k(h_k^T\omega - c_k) : \mathbf{R}^n \to \mathbf{R}, \tag{24}$$

and let

$$\phi^{(K)}(\omega_1, ..., \omega_K) = \sum_{k=1}^K \phi_k(\omega_k) : \mathbf{R}^{nK} \to \mathbf{R}, \ \ K = 1, 2, ... \tag{25}$$

Finally, let $\{x_k\}_{k=1}^K$ and $\{p_k \in \mathcal{P}^k\}_{k=1}^K$ be deterministic sequences. Then

$$\begin{array}{lrcl} (a) & \int \mathrm{e}^{-\phi^{(K)}(x_1+\xi_1,...,x_K+\xi_K)} \prod_{k=1}^K [p_k(\xi_k)d\xi_k] & \le & \prod_{k=1}^K \mathrm{risk}_{\delta_k}(\eta_k|\mathcal{P}^k) \ \ \forall(x_k\in X_1^k,\, p_k\in\mathcal{P}^k)_{k=1}^K, \\ (b) & \int \mathrm{e}^{\phi^{(K)}(x_1+\xi_1,...,x_K+\xi_K)} \prod_{k=1}^K [p_k(\xi_k)d\xi_k] & \le & \prod_{k=1}^K \mathrm{risk}_{\delta_k}(\eta_k|\mathcal{P}^k) \ \ \forall(x_k\in X_2^k,\, p_k\in\mathcal{P}^k)_{k=1}^K. \end{array}$$

For a proof, see Section A.3.

Under the premise of Proposition 2.5, consider a test T_K^η which, given K observations (5) with independent of each other $\xi_k \sim p_k \in \mathcal{P}^k$, accepts the hypothesis $H_1 : \{x_k \in X_1^k, k \le K\}$ whenever $\phi^{(K)}(\omega_1, ..., \omega_K) \ge 0$, and accepts $H_2 : \{x_k \in X_2^k, k \le K\}$ otherwise. An immediate corollary of Proposition 2.5 is as follows:

Corollary 2.2 *In the notation and under the assumptions from the premise of Proposition 2.5, combining Proposition 2.5 with Markov's inequality we obtain that the risk $\mathrm{Risk}_{\overline{\mathcal{S}}}(\mathcal{T}_K^\eta|[\mathcal{P}^k, X_1^k, X_2^k]_{k=1}^K)$ of $\mathcal{T}_K^\eta$ does not exceed the quantity $\prod_{k=1}^K \mathrm{risk}_{\delta_k}(\eta_k|\mathcal{P}^k)$.*

[1] We use the following fact (if absolutely evident facts indeed exist, this is one of them): given a vector $p = [p_1; ...; p_n]$ with entries $p_i \in [0, 1]$, and a positive integer $k \le n$, let $P_{\ge k|n}(p)$ be the probability to get $\ge k$ heads in n independent flips of a coin, with probability p_i to get a head in the i-th trial. Then $P_{\ge k|n}(p)$ is a nondecreasing function of p. Whatever "evident," this fact needs a proof, and here it is. Let $B = \{\omega \in \mathbf{R}^n : 0 \le \omega_i \le 1, i \le n\}$ be the n-dimensional cube equipped with the Lebesgue measure μ. We have $P_{\ge k|n}(p) = \mu(B_p)$, where $B_p = \{\omega \in B : \mathrm{Card}\{i : \omega_i \le p_i\} \ge k\}$. When $0 \le p_i \le p_i' \le 1$, $1 \le i \le n$, we clearly have $B_p \subset B_{p'}$, and consequently $P_{\ge k|n}(p) = \mu(B_p) \le \mu(B_{p'}) = P_{\ge k|n}(p')$, as claimed.

2.4.2 Potentials for sub-spherical families of distributions

Definition 2.3 **A.** *Given $\delta > 0$, we call a potential $\eta(s) : \mathbf{R} \to \mathbf{R}$ δ-regular, if the function*

$$H_{\delta\eta}(s) = \mathrm{e}^{-\eta(\delta - s)} + \mathrm{e}^{-\eta(\delta + s)}$$

is nondecreasing on the ray $s \geq 0$.

B. *The δ-index of the δ-regular potential η on a sub-spherical family $\mathcal{P}_\gamma$ is the quantity*

$$\epsilon_\delta(\eta|\gamma) := \int_{-\infty}^{\infty} \mathrm{e}^{-\eta(r)}\gamma(r - \delta)dr = \int_{-\infty}^{\infty} \mathrm{e}^{-\eta(\delta+s)}\gamma(s)ds = \int_0^{\infty} H_{\delta\eta}(s)\gamma(s)ds, \tag{26}$$

where the concluding equality is due to the fact that $\gamma(\cdot)$ is even.

Proposition 2.6 *Let $\mathcal{P} = \mathcal{P}_\gamma$ be a sub-spherical family of probability densities on $\mathbf{R}^n$, let $\delta \geq 0$, and let η be a δ-regular potential. Then*

$$\mathrm{risk}_\delta(\eta|\mathcal{P}_\gamma) \leq \epsilon_\delta(\eta|\gamma). \tag{27}$$

For a proof, see Section A.4.

Example 1: step potential. Assume that $\mathcal{P} = \mathcal{P}_\gamma$ is a sub-spherical family of distributions, and let $\delta > 0$, implying that $\varepsilon_\star = \varepsilon_\star(\delta|\gamma)$, as given by (19), belongs to $[0, 1/2)$ (recall that γ is an even probability density positive in a neighbourhood of the origin). We define the *step potential* as[4]

$$\eta(s) = \tfrac{1}{2}\ln\left(\frac{1 - \varepsilon_\star}{\varepsilon_\star}\right)\mathrm{sign}(s). \tag{28}$$

Taking into account that $0 \leq \varepsilon_\star < 1/2$, it is immediately seen that η is a δ-regular potential. The δ-index of the step potential satisfies

$$\epsilon_\delta(\eta|\gamma) = 2\sqrt{\varepsilon_\star(1 - \varepsilon_\star)}, \tag{29}$$

as is shown by the following computation:

$$\begin{aligned}\epsilon_\delta(\eta|\gamma) &= \int_{-\infty}^{\infty} \mathrm{e}^{-\eta(s)}\gamma(s - \delta)ds = \sqrt{\tfrac{1-\varepsilon_\star}{\varepsilon_\star}}\int_{-\infty}^{-\delta}\gamma(s)ds + \sqrt{\tfrac{\varepsilon_\star}{1-\varepsilon_\star}}\int_{-\delta}^{\infty}\gamma(s)ds \\ &= \sqrt{\tfrac{1-\varepsilon_\star}{\varepsilon_\star}}\varepsilon_\star + \sqrt{\tfrac{\varepsilon_\star}{1-\varepsilon_\star}}(1 - \varepsilon_\star) = 2\sqrt{\varepsilon_\star(1 - \varepsilon_\star)}.\end{aligned}$$

Note that in the situation of Section 2.1 with stationary K-repeated observations (4), a sub-spherical family $\mathcal{P} = \mathcal{P}_\gamma$ and non-intersecting X_1, X_2, the test $\mathcal{T}_K^\eta$ associated with the step potential η is exactly the majority test $\mathcal{T}_K^{\mathrm{maj}}$ defined in Section 2.3.2. Because the index $\epsilon_\delta(\eta|\gamma)$ of the step potential is < 1 due to $\varepsilon_\star < 1/2$, its δ-risk $\mathrm{risk}_\delta(\eta|\mathcal{P}_\gamma)$ is < 1 as well.

Example 2: ramp potential. Assume that $\mathcal{P}$ is the sub-spherical family $P_{\gamma_\mathcal{L}}$ with $\gamma_\mathcal{L}(s) = \frac{1}{2\lambda}\mathrm{e}^{-|s|/\lambda}$, $s \in \mathbf{R}$ (see the "Laplace" example in Section 2.2). Given $\delta > 0$, let us consider the *ramp potential* η_δ which minimizes the risk $\epsilon_\delta(\eta|\gamma_\mathcal{L})$:

$$\eta_\delta(s) = \tfrac{1}{2}\ln\left(\frac{\gamma_\mathcal{L}(s - \delta)}{\gamma_\mathcal{L}(s + \delta)}\right) = \begin{cases} s/\lambda & \text{for } |s| \leq \delta, \\ \frac{\delta}{\lambda}\,\mathrm{sign}(s), & \text{for } |s| > \delta. \end{cases}$$

One can easily verify that η_δ is δ-regular for any $\delta > 0$, and the corresponding δ-index satisfies

$$\epsilon_\delta(\eta_\delta|\gamma_\mathcal{L}) = \int_{-\infty}^{\infty}\sqrt{\gamma_\mathcal{L}(s - \delta)\gamma_\mathcal{L}(s + \delta)}ds = \mathrm{e}^{-\delta/\lambda}\Big(\frac{\delta}{\lambda} + 1\Big),$$

which is < 1 for all $\delta > 0$.

[4] Hereafter we put $\mathrm{sign}(s) = s/|s|$ if $s \neq 0$ and $\mathrm{sign}(0) = 0$.

Examples above show that in the situation of Section 2.1, assuming that $\mathcal{P} = \mathcal{P}_\gamma$ is a sub-spherical family, there exist potentials η with δ-indexes $\epsilon_\delta(\eta|\gamma) < 1$ and therefore, using Proposition 2.6, with risks $\mathrm{risk}_\delta(\eta|\mathcal{P}) < 1$, provided $\delta > 0$. In this situation, when solving problem $(\overline{\mathcal{S}}_K)$ with $X_1^k \equiv X_1, X_2^k \equiv X_2$, X_1 and X_2 not intersecting, and specifying $\delta > 0$ according to (18), we can decide on the hypotheses H_1 and H_2 with any desired risk $\epsilon \in (0,1)$ via semi-stationary K-repeated observations (5), provided that

$$K \geq K(\epsilon) := \left\rfloor \frac{\ln(\epsilon^{-1})}{\ln(\mathrm{risk}_\delta(\eta|\mathcal{P})^{-1})} \right\lfloor, \tag{30}$$

see Corollary 2.2 (here $\lfloor x \rfloor$ stands for the smallest integer larger or equal to x).

2.4.3 Potentials for the sub-Gaussian family

Aside from sub-spherical families, there are other families of probability distributions allowing, in the case of $X_1 \cap X_2 = \emptyset$, for potentials with risks < 1 and thus for tests with an arbitrarily low risk, provided stationary or semi-stationary K-repeated observations with properly selected K are available. The simplest family of this type is the family $\mathcal{P}_{\mathrm{sG}}^n$ of sub-Gaussian probability densities $p(\cdot)$, with parameters 0 and I_n, that is, probability densities p on $\mathbf{R}^n$ such that

$$\int \mathrm{e}^{h^T\xi} p(\xi) d\xi \leq \mathrm{e}^{\frac{1}{2}h^T h} \quad \forall h \in \mathbf{R}^n. \tag{31}$$

Assuming that $\mathcal{P} = \mathcal{P}_{\mathrm{sG}}$ and given $\delta > 0$, let us put

$$\eta_{\mathrm{sG},\delta}(s) = \delta s. \tag{32}$$

Proposition 2.7 *Whenever $\delta \geq 0$, we have*

$$\mathrm{risk}_\delta(\eta_{\mathrm{sG},\delta}|\mathcal{P}_{\mathrm{sG}}) \leq \mathrm{e}^{-\delta^2/2}. \tag{33}$$

Proposition 2.7 is a special case of Proposition 3.3 in [13]; to make the presentation self-contained, we provide its proof in Section A.5.

Note that if $\mathcal{P}$ is the sub-spherical family $\mathcal{P}_{\gamma_\mathcal{G}}$ with $\gamma_\mathcal{G}(s) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-s^2/2}$, $s \in \mathbf{R}$ (recall that this family contains, for instance, Gaussian distributions with zero mean and covariance matrix $\preceq I_n$, cf. Section 2.2.1) the potential $\eta_{\mathrm{sG},\delta}(s) = \frac{1}{2}\ln\left(\frac{\gamma_\mathcal{G}(s-\delta)}{\gamma_\mathcal{G}(s+\delta)}\right) = \delta s$ minimizes the δ-index over $\mathcal{P}_{\gamma_\mathcal{G}}$ with $\epsilon_\delta(\eta_{\mathrm{sG},\delta}|\gamma_\mathcal{G}) = \mathrm{e}^{-\delta^2/2}$.

2.4.4 "Majority of means" tests

Let now $X_1 \subset \mathbf{R}^n$ and $X_2 \subset \mathbf{R}^n$ be closed convex nonempty sets, one of the sets being bounded, and such that $X_1 \cap X_2 = \emptyset$. In this situation, given $\epsilon \in (0,1)$ and K-repeated stationary observations ω^K, the potential-based tests developed in Sections 2.4.1 – 2.4.3 allow to decide on H_1, H_2 with risk $\leq \epsilon$, where K grows logarithmically with ϵ^{-1}. For instance, in the case of a sub-Gaussian family of distributions, the corresponding test attains the risk ϵ provided that $K \geq K(\epsilon) = O\left(\ln(\epsilon^{-1})/\delta^2\right)$ (cf. (33)), where $2\delta > 0$ is the Euclidean distance between X_1 and X_2, see (16) – (18). On the other hand, these tests rely upon the "Cramer-type" Definition 2.2 of the risk of the potential, and thus assume control of the exponential moment of $\eta(\xi)$. In this section we present a different testing procedure, which only uses second order characteristics of the potential (mean and variance), and yet allows to achieve arbitrarily low risks of testing for essentially the same sizes of observation sample. We describe this modification in the simplest case of stationary repeated observations, generalizations to more general settings (e.g., that of Proposition 2.5) being straightforward.

Now, let x_*^1, x_*^2 and h_* be associated with X_1 and X_2 via (16) and (17), and let $\mathcal{P}$ be a family of probability distributions on $\mathbf{R}^n$. We suppose that a potential $\eta(\cdot)$ and $c \in \mathbf{R}$ are such that for some $\varrho > 0$ and all $p \in \mathcal{P}$,

$$\begin{array}{ll} (a) & \mathbf{E}_{\xi\sim p}\{\eta(h_*^T(x_*^1+\xi)+c)\} - \mathbf{E}_{\xi\sim p}\{\eta(h_*^T(x_*^2+\xi)+c)\} \geq \varrho, \\ (b) & \mathbf{Var}_{\xi\sim p}\{\eta(h_*^T(x+\xi)+c)\} \leq 1 \end{array} \tag{34}$$

for all $x \in X_1 \cup X_2$.[5]

[5]Here and below **Var** denotes the "usual" variance: for a probability density p on $\mathbf{R}^n$ and $f : \mathbf{R}^n \to \mathbf{R}$, $\mathbf{Var}_{\xi\sim p}\{f(\xi)\} = \mathbf{E}_{\xi\sim p}\{f^2(\xi)\} - [\mathbf{E}_{\xi\sim p}\{f(\xi)\}]^2$.

Example. Let $\mathcal{P}$ be a family of zero-mean distributions on $\mathbf{R}^n$ with covariance matrix $\preceq \sigma^2 I_n$:

$$\int \xi p(\xi)d\xi = 0, \quad \int \xi\xi^T p(\xi)d\xi \preceq I_n, \quad \forall p \in \mathcal{P}.$$

For the linear potential $\eta(t) = t$ with $c = 0$ we clearly have

$$\begin{array}{l}\mathbf{E}_{\xi\sim p}\{\eta(h_*^T(x_*^1+\xi)+c)\} - \mathbf{E}_{\xi\sim p}\{\eta(h_*^T(x_*^2+\xi)+c)\} = h_*^T(x_*^1 - x_*^2) = 2\delta =: \varrho, \quad \forall p \in \mathcal{P} \\ \mathbf{Var}_{\xi\sim p}\{\eta(h_*^T(x+\xi)+c)\} = \mathbf{E}_{\xi\sim p}\{(h_*^T\xi)^2\} \le 1 \ \ \forall (x \in X_1 \cup X_2,\ p \in \mathcal{P})\end{array}$$

(recall that $\delta = \frac{1}{2}\|x_*^1 - x_*^2\|_2$, and $\|h_*\|_2 = 1$).

Now, given K-repeated stationary observations $\omega^K = [x+\xi_1, ..., x+\xi_K]$, and $\epsilon_1, \epsilon_2 \in (0,1)$, we consider the inference problem $(\mathcal{S}_K)$ of deciding via the observation ω^K whether $x \in X_1$ (hypothesis H_1) or $x \in X_2$ (hypothesis H_2). Our objective is to build the test $\mathcal{T}_K$ for $\mathcal{S}^K$ such that, uniformly over $x \in X_1 \cup X_2$ and $p \in \mathcal{P}$, the probability of wrongly rejecting H_1 (accepting H_2) is $\le \epsilon_1$, and the probability of wrongly rejecting H_2 (accepting H_1) is $\le \epsilon_2$. and we want to attain this goal using the smallest possible size K of observation sample ω^K.

For the sake of definiteness, assume that $\epsilon_1 \le \epsilon_2$. We denote $\kappa = \frac{1}{2}\left(1 - \frac{\ln(\epsilon_2^{-1})}{\ln(\epsilon_1^{-1})}\right)$ (note that $0 \le \kappa < \frac{1}{2}$). Let now $m = \lfloor 4\mathrm{e}(\mathrm{e}^{-\kappa}+1)^2\varrho^{-2}\rfloor$, and let

$$\psi_j(\omega^K) = \frac{1}{m}\sum_{i=(j-1)m+1}^{mj} \eta(h_*^T\omega_i + c), \quad j = 1, 2, ...$$

Denote

$$c_* = \mathbf{E}_{\xi\sim p}\{\eta(h_*^T(x_*^1+\xi)+c)\} - \frac{\mathrm{e}^\kappa\varrho}{1+\mathrm{e}^\kappa} \ge \mathbf{E}_{\xi\sim p}\{\eta(h_*^T(x_*^2+\xi)+c)\} + \frac{\varrho}{1+\mathrm{e}^\kappa}.$$

The *K-observation majority of means test* $\mathcal{T}_K^{\mathrm{mm}}$ for $\mathcal{S}_K$ is as follows: given

$$K := Jm \ge \lfloor 2\ln(\epsilon_1^{-1})\rfloor m \tag{35}$$

observations ω_k, $k = 1, ..., K$, $\mathcal{T}_K^{\mathrm{mm}}$ accepts the hypothesis H_1 when $\psi_j(\omega^K) \ge c_*$ for at least $J/2$ values of j, and accepts the hypothesis H_2 otherwise.

Proposition 2.8 *In the just described situation, the risks of the test $\mathcal{T}_K^{\mathrm{mm}}$ meet the problem specifications, namely,*

$$\mathrm{Risk}_{1\mathcal{S}}(\mathcal{T}_K|\mathcal{P}, X_1, X_2) \le \epsilon_1, \quad \mathrm{Risk}_{2\mathcal{S}}(\mathcal{T}_K|\mathcal{P}, X_1, X_2) \le \epsilon_2.$$

For a proof, see Section A.6.

Let us consider the test $\mathcal{T}_K^{\mathrm{mm}}$ using the linear potential with $c = 0$. Under the premise of the proposition, i.e., in the situation where the noise covariance matrix is $\preceq I_n$ and the distance between the sets X_1 and X_2 is $\ge 2\delta$, the size of the stationary K-repeated observation sufficient for the test to satisfy the risk specifications is

$$\lfloor \mathrm{e}(\mathrm{e}^{-\kappa}+1)^2\delta^{-2}\rfloor\, \lfloor 2\ln(\epsilon_1^{-1})\rfloor = O\big(\delta^{-2}\ln(\epsilon_1^{-1})\big).$$

Note that when $\epsilon_1 = \epsilon_2 = \epsilon$, the above number of observations sufficient to decide on H_1, H_2 with risk ϵ is within absolute constant factor of the number of observations, as given by (30) and (33), needed for the same purpose in the case when $\mathcal{P} = \mathcal{P}_{\mathrm{sG}}$.

3 Sequential detection via Euclidean separation

3.1 Motivating example

We start introducing a motivating example which will help understand the general setting discussed in Sections 3.2, 3.3, 3.4 and which will be used in Section 4.3 to test our methodology.

Consider the simple time series model

$$\begin{array}{rcl} y_k & = & \alpha_k + \zeta_k \\ \alpha_k & = & \alpha_{k-1} + \eta_k + u_k \end{array}, \ k = 1, 2, ..., d, \tag{36}$$

where

- y_k is the observation at time k, and $u = [u_1; ...; u_d] \in \mathbf{R}^d$ is the deterministic input,
- $\zeta = [\zeta_1; ...; \zeta_d]$ is zero mean d-dimensional Gaussian random vector with unknown covariance matrix known to be $\preceq \sigma^2 I_d$;
- $\eta = [\eta_1; ...; \eta_d]$ is independent of ζ d-dimensional random vector obeying multivariate Student distribution with ν degrees of freedom and matrix parameter $\preceq I_d$. Here ν is a positive integer or $+\infty$, with $\nu = \infty$ interpreted as the fact that η is a zero mean Gaussian random vector with covariance matrix $\preceq I_d$.

We intend to decide from observations $y_1, ..., y_d$ on the nuisance hypothesis $u = 0$ vs. signal alternative

$$u \in \bigcup_{\substack{1 \le i \le d, \\ 0 < \rho < R}} [U_i^+(\rho) \cup U_i^-(\rho)], \tag{37}$$

where $R \in (0, \infty)$ is a parameter, $U_i^-(\rho) = -U_i^+(\rho)$, and $[U_i^+(\rho) \cup U_i^-(\rho)]$ is the set of "signal inputs of shape i and magnitude $\ge \rho$." We consider two cases:

- *pulse signal inputs:* $U_i^+(\rho) = \{u \in \mathbf{R}^d : u_k = 0, k \ne i, \rho \le u_i \le R\}$, $1 \le i \le d$;
- *step signal inputs:* $U_i^+(\rho) = \{u \in \mathbf{R}^d : u_k = 0, k < i, \rho \le u_i = u_{i+1} = ... = u_d \le R\}$, $1 \le i \le d$.

To make the notation consistent with the one used in the general setting described in Section 3.2, we set $U_{2i-1}(\rho) = U_i^+(\rho)$ and $U_{2i}(\rho) = U_i^-(\rho)$, thus getting $N = 2d$ parametric families of signal inputs and we refer to signal input $u \in U_j(\rho)$ as to signal input of shape j and magnitude $\ge \rho$.

Control parameters are $R > 0$ appearing in (37) and tolerance $\epsilon \in (0, 1/2)$ responsible for the risks of our decision rules.

Our goal is to find decision rules $\mathcal{T}_k$ and positive reals $\rho_{kj} \in (0, R]$, $1 \le k \le d$, $1 \le j \le N$, such that

- $\mathcal{T}_k$ makes a decision given observation $y^k = [y_1; ...; y_k]$, and this decision, depending on y^k, is either "signal conclusion," or "nuisance conclusion" (exactly one of them). In the case of signal conclusion at step k, the inference procedure is terminated at this step, otherwise we pass to time $k+1$ (when $k < d$) or terminate (when $k = d$).
- The following risk specifications are met:
  - When the nuisance hypothesis is true (i.e., $u = 0$), the probability of terminating with signal conclusion (false alarm) somewhere on the time horizon $1, ..., d$ is $\le \epsilon$.
  - For every $k \le d$ and every $j \le N$, if the input u underlying our observation belongs to $U_j(\rho)$ with $\rho_{kj} \le \rho < R$, the probability of signal conclusion somewhere on the time horizon $1, ..., k$ should be at least $1 - \epsilon$.

We intend to achieve this goal with as small ρ_{kj} as possible.

We now explain how to cast the problem we have just described as the more general detection problem discussed in Sections 3.2, 3.3, 3.4. To this end, we need to build a new observation scheme. More precisely, we have

$$y^k = \alpha_0 \underbrace{[1; ...; 1]}_{k} + B_k \eta + C_k \zeta + B_k u$$

for some known matrices B_k and C_k. Let F_k be $d \times (k-1)$ matrix with columns forming an orthonormal basis of the $(k-1)$-dimensional subspace of $\mathbf{R}^k$ comprised of vectors with zero mean. Let us set

$$z^k := F_k^T y^k = D_k \eta + E_k \zeta + D_k u \tag{38}$$

where $(k-1) \times d$ matrices D_k and E_k are given by $D_k = F_k^T B_k$, $E_k = F_k^T C_k$, and, as is immediately seen, have rank $k-1$. We treat z^k, rather than y^k, as an intermediate observation at time k.

We then find a positive definite matrix Θ_k such that

$$\Theta_k^2 \succeq D_k D_k^T \ \& \ \Theta_k^2 \succeq E_k E_k^T.$$

Finally, the observation ω^k at time k, is the vector (cf. (54))

$$\omega^k = \Theta_k^{-1} z^k = \underbrace{[\Theta_k^{-1} D_k]}_{A_k} u + \underbrace{\Theta_k^{-1} [D_k \eta + E_k \zeta]}_{\xi_k}. \tag{39}$$

We have come to observation scheme (39) which can be handled by the change detection procedure we are about to describe in Sections 3.2, 3.3, 3.4. In particular, by Proposition 2.3, ξ_k has a symmetric density $p_k(\cdot) \in \mathcal{P}_\gamma^{k-1}$ with $\gamma = \gamma_{\mathcal{S}} \star \gamma_\sigma$, where $\gamma_{\mathcal{S}}$ is the density of the standard univariate Student's t_ν distribution with ν degrees of freedom, and γ_σ is the density of $\mathcal{N}(0, \sigma^2)$.

3.2 Situation and goal

In this Section our objective is to make decisions about an unknown vector $u \in \mathbf{R}^{n_u}$ representing inputs of a linear system in the situation where the information about u is acquired sequentially, and the goal is to decide whether the input observed so far is a nuisance or is "meaningful." Our modeling methodology goes back to [5]. Specifically, suppose that the observation $\omega^k \in \mathbf{R}^{m_k}$ available at step $k = 1, ..., K$, is

$$\omega^k = A_k u + \xi_k, \tag{40}$$

where u is the unknown system input, $\xi_k \in \mathbf{R}^{m_k}$ are random noises, and $A_k \in \mathbf{R}^{m_k \times n_u}$ are known matrices. Throughout this section we suppose that we are given

1. a family $\mathcal{P}$ of distributions on $\mathbf{R}^{n_\xi}$ and matrices $Z_k \in \mathbf{R}^{m_k \times n_\xi}$ such that Z_k is of rank m_k and
$$\xi_k = Z_k \xi, \ k = 1, ..., K, \tag{41}$$
where ξ obeys some (possibly unknown) distribution $p_\xi \in \mathcal{P}$;
2. a convex compact set $U_{\rm inp} \subset \mathbf{R}^{n_u}$ of *admissible inputs*;
3. a convex compact *nuisance set* $V_{\rm nui} \subset \operatorname{int} U_{\rm inp}$ such that $0 \in V_{\rm nui}$;
4. N convex and closed *"activation" sets* $W_j \subset \mathbf{R}^{n_u}$ such that $0 \notin W_j$ and
$$w \in W_j, \rho \geq 1 \Rightarrow \rho w \in W_j; \tag{42}$$
5. N nonempty convex compact "drag sets" $V_{\rm drg}^j$ such that $0 \in V_{\rm drg}^j$.

For the example of Section 3.1, we have $\xi = [\eta; \zeta]$, $Z_k[\eta; \zeta] = \Theta_k^{-1} [D_k \eta + E_k \zeta]$, $U_{\rm inp} = \{u \in \mathbb{R}^d : 0 \leq u_i \leq R\}$, $V_{\rm nui} = V_{\rm drg}^j = \{0\}$, and

- for *pulse signal inputs:* $W_{2i-1} = \{u \in \mathbf{R}^d : u_k = 0, k \neq i, 1 \leq u_i\}$, $W_{2i} = -W_{2i-1}$, $1 \leq i \leq d$;
- for *step signal inputs:* $W_{2i-1} = \{u \in \mathbf{R}^d : u_k = 0, k < i, 1 \leq u_i = u_{i+1} = ... = u_d\}$, $W_{2i} = -W_{2i-1}$, $1 \leq i \leq d$.

We call an input $u \in V_{\rm nui}$ a *nuisance*, and a vector z of the form $z = \rho w$, with $w \in W_j$ and $\rho > 0$, an *activation of shape j and magnitude $\geq \rho$*. Note that if $0 < \rho' \leq \rho$ and z is an activation of shape j and magnitude $\geq \rho$, then, as it should be, z is an activation of shape j and magnitude $\geq \rho'$, due to $z = \rho' w'$, $w' = (\rho/\rho') w$, and $w' \in W_j$ due to $0 < \rho' \leq \rho$ and (42).

We call input u a *signal of shape j and magnitude $> \rho$ (or $\geq \rho > 0$)*, if $u \in U_{\rm inp}$ and

$$u = v + \rho' w$$

with $v \in V_{\rm drg}^j$, $w \in W_j$, and $\rho' > \rho$ (resp. $\rho' \geq \rho$), and denote by

$$U_j(\rho) = \{u = v + \rho w : u \in U_{\rm inp}, v \in V_{\rm drg}^j, w \in W_j\}$$

the set of all signal inputs of shape j and magnitude $\geq \rho$.

Our goal is to decide via observations ω^k, $k \le K$, on the nuisance hypothesis "the input u to (40) is a nuisance" (i.e., $u \in V_{\rm nui}$) vs. the signal alternative "the input u to (40) is a signal of some shape and (positive) magnitude." Note that in this case, under the signal alternative the input u belongs to a nonconvex set expressed as a union of convex sets. More precisely, we assume that we are given tolerances

$$\{\epsilon_k \in (0,1/2),\, 1 \le k \le K\},\ \{\epsilon_{kj} \in (0,1/2), 1 \le k \le K, 1 \le j \le N\}$$

and want to design a sequence of decision rules $\{\mathcal{T}_k : 1 \le k \le K\}$ along with thresholds $\rho_{kj} > 0$, $1 \le k \le K$, $1 \le j \le N$ with the following properties: for every $k \le K$,

- rule $\mathcal{T}_k$ makes a decision based on observation ω^k, and this decision is either to accept the null hypothesis or to accept the signal one (but not both);
- if the input is a nuisance, the probability for $\mathcal{T}_k$ to accept the signal alternative is $\le \epsilon_k$;
- if, for some j, the input is a signal of shape j and magnitude $> \rho_{kj}$, the probability for $\mathcal{T}_k$ to accept the nuisance hypothesis is at most ϵ_{kj}.

Given tolerances ϵ_k, ϵ_{kj}, we would like to achieve the outlined goal with thresholds ρ_{kj} as small as possible.

3.3 Tests $\mathcal{T}_k$, Scheme I.

3.3.1 Assumptions on the distribution of noise ξ

Throughout Section 3.3, we make the following assumption on the family $\mathcal{P}$ of probability densities of random disturbance ξ in (41):

Assumption A1 *For every $k \le K$ we can point out a parametric family $\mathcal{H}_k = \{\eta_\delta^k(\cdot) : \delta \ge 0\}$ of potentials and a continuous nonincreasing function $\mathcal{R}_k(\delta) : \mathbf{R}_+ \to (0,1]$ such that $\mathcal{R}_k(0) = 1$ and*

$$\mathrm{risk}_\delta(\eta_\delta^k|\mathcal{P}^k) \le \mathcal{R}_k(\delta)\ \forall \delta \ge 0, \tag{43}$$

where $\mathcal{P}^k$ is the family of probability densities of random vectors $\xi_k = Z_k\xi$ with $\xi \sim p_\xi \in \mathcal{P}$.

Note that Assumption A1 indeed holds in the situations of our primary interest. Specifically, assume that $Z_k Z_k^T \preceq I_{m_k}$.[6] Then

- when $\mathcal{P} = \mathcal{P}_{\rm sG}^{n_\xi}$ is the family of sub-Gaussian distributions on $\mathbf{R}^{n_\xi}$, with parameters $0, I_{n_\xi}$, then $\mathcal{P}^k$ belongs to the family $\mathcal{P}_{\rm sG}^{m_k}$ of sub-Gaussian, with parameters $0, I_{m_k}$, distributions on $\mathbf{R}^{m_k}$. By Proposition 2.7, (43) is ensured by the choice

$$\mathcal{H}_k = \{\eta_\delta^k(s) = \delta s, \delta \ge 0\},\ \ \mathcal{R}_k(\delta) = \mathrm{e}^{-\delta^2/2};$$

- when $\mathcal{P} = \mathcal{P}_\gamma$ is a sub-spherical family of distributions on $\mathbf{R}^{n_\xi}$, (41) combines with Proposition 2.1 to ensure that the family $\mathcal{P}^k$ is contained in the sub-spherical family of distributions $\mathcal{P}_\gamma^{m_k}$. Invoking the example of the step potential from Section 2.4.2, (43) is ensured by setting

$$\mathcal{H}_k = \left\{\eta_\delta^k(s) = \tfrac{1}{2}\ln\left(\tfrac{1-\varepsilon_\star(\delta)}{\varepsilon_\star(\delta)}\right)\mathrm{sign}(s), \delta \ge 0\right\},\ \ \mathcal{R}_k(\delta) = 2\sqrt{\varepsilon_\star(\delta)(1-\varepsilon_\star(\delta))},$$
$$\varepsilon_\star(\delta) = \int_\delta^\infty \gamma(s)ds.$$

3.3.2 Building decision rules

For every $k \le K$ and every $j \le N$, consider the parametric convex optimization problem

$$\mathrm{Opt}_{kj}(\rho) = \min_{v',v,w}\left\{\tfrac{1}{2}\|A_k(v' - [v+\rho w])\|_2 : v' \in V_{\rm nui}, v \in V_{\rm drg}^j, w \in W_j, v+\rho w \in U_{\rm inp}\right\} \tag{$P_{kj}[\rho]$}$$

where the parameter ρ is positive.

From our assumptions on $V_{\rm nui}$, $U_{\rm inp}$, $V_{\rm drg}^j$ and W_j it immediately follows that

[6] this always can be achieved by appropriate scaling of observations (40).

1. the set Δ_j of those $\rho > 0$ for which $(P_{kj}[\rho])$ is feasible, is a half-open segment $(0, R_j]$ with $0 < R_j < \infty$, and
2. when $\rho \in \Delta_j$, problem $(P_{kj}[\rho])$ is solvable, and $\mathrm{Opt}_{kj}(\rho)$ is a real-valued continuous nondecreasing function on Δ_j such that $\lim_{\rho\to+0} \mathrm{Opt}_{kj}(\rho) = 0$.

Given $k \le K$, we specify the test $\mathcal{T}_k$ as follows.

1. We compute the quantities R_j. [7]
2. We select somehow a (perhaps, empty) set $J_k \subset \{1, 2, ..., N\}$ and reals $\rho_{kj} \in (0, R_j]$, $j \in J_k$, in such a way that
   (a) $\rho_{kj} = R_j$ when $j \notin J_k$,
   (b) we have
   $$\sum_{j\in J_k} \epsilon_{kj}^{-1} \mathcal{R}_k^2(\mathrm{Opt}_{kj}(\rho_{kj})) \le \epsilon_k. \tag{44}$$
   Note that (44) implies that $\mathrm{Opt}_{kj}(\rho_{kj}) > 0$ for $j \in J_k$, otherwise the left hand side in (44) is at least 1 due to $\mathcal{R}_k(0) = 1$, and we have assumed that $\epsilon_k < 1/2$.
3. If $J_k = \emptyset$, $\mathcal{T}_k$ accepts the nuisance hypothesis. When $J_k \neq \emptyset$, we act as follows:
   (a) for $j \in J_k$, we solve the optimization problem $(P_{kj}[\rho_{kj}])$ and denote by $(v'_{kj}; v_{kj}, w_{kj})$ an optimal solution to the problem. We set
   $$\begin{array}{c} \delta_{kj} = \mathrm{Opt}_{kj}(\rho_{kj}), \quad u_{kj} = v_{kj} + \rho_{kj} w_{kj}, \quad h_{kj} = \frac{A_k[v'_{kj} - u_{kj}]}{\|A_k[v'_{kj} - u_{kj}]\|_2}, \\ c_{kj} = \frac{1}{2} h_{kj}^T A_k[v'_{kj} + u_{kj}], \quad \phi_{kj}(\omega^k) = \eta^k_{\delta_{kj}}(h_{kj}^T \omega^k - c_{kj}), \end{array} \tag{45}$$
   with $\eta^k_\delta(\cdot)$ given by Assumption A1, and define α_{kj} from the relation
   $$\mathrm{e}^{\alpha_{kj}} = \epsilon_{kj}/\mathcal{R}_k(\delta_{kj}). \tag{46}$$
   (b) Finally, given observation ω^k, the rule $\mathcal{T}_k$ accepts the nuisance hypothesis if $\phi_{kj}(\omega^k) + \alpha_{kj} \ge 0$ for all $j \in J_k$, and accepts the signal hypothesis otherwise.

3.3.3 Performance analysis

We now check that the just defined decision rules meet the goal stated in Section 3.2. Let us fix $k \le K$ and $j \in J_k$. Given feasible input u, let P_u^k be the distribution of observation ω^k, the input being u. By construction and due to Proposition 2.5, applied to observations $\omega^k = A_k u + \xi_k$ and $X_1 = A_k V_{\mathrm{nui}}$, $X_2 = A_k U_j(\rho_{kj})$, we have

$$\begin{array}{rcll} \int \mathrm{e}^{-\phi_{kj}(\omega^k)} P_u^k(d\omega^k) & \le & \mathcal{R}_k(\delta_{kj}), & \text{when } u \text{ is a nuisance,} \\ \int \mathrm{e}^{\phi_{kj}(\omega^k)} P_u^k(d\omega^k) & \le & \mathcal{R}_k(\delta_{kj}), & \text{when } u \text{ is a signal of shape } j \text{ and magnitude} \ge \rho_{kj}. \end{array} \tag{47}$$

Assume first that the input u is a nuisance, and let us upper-bound the P_u^k-probability of rejecting the nuisance hypothesis at step k. This takes place only when $\phi_{kj}(\omega^k) + \alpha_{kj} < 0$ for some $j \in J_k$, and, by the first relation in (47), for a given $j \in J_k$ the P_u^k-probability of this event is at most $\mathrm{e}^{-\alpha_{kj}} \mathcal{R}_k(\delta_{kj}) = \mathcal{R}_k^2(\delta_{kj})/\epsilon_{kj}$ where the last equality is due to (46). As a result, the P_u^k-probability to reject the nuisance hypothesis at step k is at most

$$\sum_{j\in J_k} \epsilon_{kj}^{-1} \mathcal{R}_k^2(\delta_{kj}) = \sum_{j\in J_k} \epsilon_{kj}^{-1} \mathcal{R}_k^2(\mathrm{Opt}_{kj}(\rho_{kj})) \le \epsilon_k,$$

where the concluding inequality is due to (44).

Now, let the input u be a signal of shape j and magnitude $> \rho_{kj}$. Due to $\rho_{kj} = R_j$ for $j \notin J_k$, we have $j \in J_k$. Let us upper-bound the P_u^k-probability of the nuisance conclusion at step k. By construction, the latter

[7] The simplest way to identify R_j is to run bisection in ρ on a large initial range of ρ in order to find the largest ρ for which $(P_{kj}[\rho])$ is feasible.

occurs only when $\phi_{kj}(\omega^k)+\alpha_{kj}\geq 0$, and, by the second relation in (47), this can happen with P_u^k-probability at most $e^{\alpha_{kj}}\mathcal{R}_k(\delta_{kj})=\epsilon_{kj}$, where the concluding equality is due to (46).

The bottom line is that the probability of false alarm at step k (a signal conclusion when the input is nuisance) is $\leq\epsilon_k$, and the probability of ρ_{kj}-miss (the probability to make a nuisance conclusion when the input is a signal of shape j and magnitude $>\rho_{kj}$) is at most ϵ_{kj}.

3.4 Decision rules $\mathcal{T}_k$, Scheme II

Throughout Section 3.4, we make the following

> **Assumption A:** The family $\mathcal{P}$ of probability densities of the disturbance ξ in (41) is such that for all $k\leq K$, probability densities of noises ξ_k belong to sub-spherical families $\mathcal{P}_\gamma^{m_k}$ with some common γ.

Note this assumption takes place, e.g., when $\mathcal{P}$ is a sub-spherical family $\mathcal{P}_\gamma^{n_\xi}$ and $Z_kZ_k^T\preceq I_{m_k}$, $k\leq K$, see Proposition 2.1.

We put (cf. (19))

$$\varepsilon_\star(\delta|\gamma)=\int_\delta^\infty\gamma(s)ds, \tag{48}$$

3.4.1 Building the decision rules

Given $k\leq K$, we specify $\mathcal{T}_k$ as follows.

1. Exactly as in Scheme I, for every $j\leq N$, we consider the parametric convex optimization problem $(P_{kj}[\rho])$ with positive ρ, and compute R_j, the largest ρ for which the problem is feasible.

2. We select somehow a (perhaps, empty) set $J_k\subset\{1,2,...,N\}$ and reals $\rho_{kj}\in(0,R_j]$, $1\leq j\leq N$, and $\alpha_{\chi kj}$, $\chi=1,2$, $j\in J_k$ in such a way that $\rho_{kj}=R_j$ when $j\notin J_k$, and we have

$$\begin{array}{lrcl}(a) & \varepsilon_\star(\alpha_{2kj}|\gamma) & \leq & \epsilon_{kj},\ \ \forall j\in J_k,\\ (b) & \sum\limits_{j\in J_k}\varepsilon_\star(\alpha_{1kj}|\gamma) & \leq & \epsilon_k,\\ (c) & \alpha_{1kj}+\alpha_{2kj} & \leq & 2\delta_{kj},\ \ \delta_{kj}=\mathrm{Opt}_{kj}(\rho_{kj}),\ \ \forall j\in J_k.\end{array} \tag{49}$$

   Note that for $j\in J_k$ we have $\alpha_{1kj}>0$ and $\alpha_{2kj}>0$ (by (49.a-b) combined with $\epsilon_{kj}<1/2$, $\epsilon_k<1/2$; recall that $\varepsilon_\star(s|\gamma)\geq 1/2$ when $s\leq 0$). As a result, (49.c) implies that $\delta_{kj}>0$ whenever $j\in J_k$.

3. If $J_k=\emptyset$, $\mathcal{T}_k$ accepts the nuisance hypothesis. When $J_k\neq\emptyset$, we act as follows:

   (a) for $j\in J_k$, we solve the optimization problem $(P_{kj}[\rho_{kj}])$ and denote by $(v'_{kj};v_{kj},w_{kj})$ an optimal solution to the problem. Similarly to (45), we set

$$\begin{array}{c}u_{kj}=v_{kj}+\rho_{kj}w_{kj},\quad h_{kj}=\frac{A_k[v'_{kj}-u_{kj}]}{\|A_k[v'_{kj}-u_{kj}]\|_2},\\ c_{kj}=\frac{1}{2}h_{kj}^TA_k[v'_{kj}+u_{kj}],\quad \phi_{kj}(\omega^k)=h_{kj}^T\omega^k-c_{kj}.\end{array} \tag{50}$$

   (b) Finally, given observation ω^k, the rule $\mathcal{T}_k$ accepts the nuisance hypothesis if $\phi_{kj}(\omega_k)\geq\frac{1}{2}(\alpha_{2kj}-\alpha_{1kj})$ for all $j\in J_k$, and accepts the signal hypothesis otherwise.

3.4.2 Performance analysis

Let $k\leq K$ be fixed, and let P_u^k be the probability distribution of observation ω^k, the input being u. Taking into account (49.c) and applying Proposition 2.4 with $\alpha_\chi=\alpha_{\chi kj}$, $\delta=\delta_{kj}$ and $X_1=A_kV_{\rm nui}$, $X_2=A_kU_j(\rho_{kj})$, we obtain

$$\begin{array}{lrcl}(a) & P_u^k\left\{\phi_{kj}(\omega^k)<\frac{1}{2}(\alpha_{2kj}-\alpha_{1kj})\right\} & \leq & \varepsilon_\star(\alpha_{1kj}|\gamma),\quad \text{if } u\in V_{\rm nui},\\ (b) & P_u^k\left\{\phi_{kj}(\omega^k)\geq\frac{1}{2}(\alpha_{2kj}-\alpha_{1kj})\right\} & \leq & \varepsilon_\star(\alpha_{2kj}|\gamma),\quad \text{if } u\in U_j(\rho_{kj}).\end{array} \tag{51}$$

Assume first that the input u is a nuisance, and let us upper-bound the P_u^k-probability of rejecting the nuisance hypothesis at step k. This rejection implies that $\phi_{kj}(\omega^k) < \frac{1}{2}(\alpha_{2kj} - \alpha_{1kj})$ for some $j \in J_k$, and by (51.a) the P_u^k-probability of this event for a given $j \in J_k$ is at most $\varepsilon_\star(\alpha_{1kj}|\gamma)$. As a result, the P_u^k-probability of signal conclusion at step k when u is a nuisance is at most

$$\sum_{j \in J_k} \varepsilon_\star(\alpha_{1kj}|\gamma) \le \epsilon_k,$$

where the inequality is due to (49.b).

Now let the input u be a signal of shape j and magnitude $> \rho_{kj}$, i.e., $u \in U_j(\rho)$ with $\rho > \rho_{kj}$. Due to $\rho_{kj} = R_j$ for $j \notin J_k$ this means that $j \in J_k$. Let us upper-bound the P_u^k-probability of a nuisance conclusion at step k. The nuisance hypothesis is not rejected only when $\phi_{kj}(\omega^k) \ge \frac{1}{2}(\alpha_{2kj} - \alpha_{1kj})$, and by (51.b) the P_u^k-probability of the latter event does not exceed $\varepsilon_\star(\alpha_{2kj}|\gamma) \le \epsilon_{kj}$ (recall that by definition of sets $U_j(\rho)$, $u \in U_j(\rho)$ with $\rho > \rho_{kj}$ implies that $u \in U_j(\rho_{kj})$), where the concluding inequality is due to (49.a).

The bottom line is that the P_u^k-probability of false alarm at step k (rejecting the nuisance hypothesis when it is true) is $\le \epsilon_k$, and the P_u^k-probability of ρ_{kj}-miss (making a nuisance conclusion when the input is a signal of shape j and magnitude $> \rho_{kj}$) is at most ϵ_{kj}.

4 Application: change detection in linear dynamical system

4.1 Problem statement

We consider the change detection problem as follows.

1. We are given a discrete time linear time invariant system

$$\begin{array}{rcl} x_t & = & P_t x_{t-1} + Q_t u + R_t \xi, \\ y_t & = & C_t x + D_t u + S_t \xi, \quad t = 1, ..., d, \end{array} \tag{52}$$

   where

   - $x = [x_0; x_1; ...; x_d]$, $x_t \in \mathbf{R}^{n_x}$ is the state trajectory, $u \in \mathbf{R}^{n_u}$ is the input, $y_t \in \mathbf{R}^{n_y}$ is the output at time t,
   - $\xi \in \mathbf{R}^{n_\xi}$ is a random disturbance with (unknown) distribution P,
   - P_t,...,S_t, $1 \le t \le d$, are known matrices of appropriate sizes.

   Given $\tau \le d$, we set $y^\tau = [y_1; ...; y_\tau]$. We denote $E_x = \underbrace{\mathbf{R}^{n_x} \times ... \times \mathbf{R}^{n_x}}_{d}$ and similarly for E_u.

2. We are given sets $U_{\rm inp} \subset E_u$ (admissible inputs), $V_{\rm nui} \subset E_u$ (nuisances), $V_{\rm drg}^j$ (drags), $W_j \subset E_u$ (activations of shape j and magnitude ≥ 1), $1 \le j \le N$, meeting the requirements of Section 3.2. These sets, exactly as in Section 3.2, give rise to the notions of admissible and nuisance inputs to (52), same as signal inputs of shape j and magnitude $\ge \rho$.

3. The distribution P of disturbance ξ is known to belong to a given family $\mathcal{P}$ of probability densities on $\mathbf{R}^{n_\xi}$. We are also given tolerances $\epsilon_t \in (0, 1/2)$, $\epsilon_{tj} \in (0, 1/2)$, $1 \le t \le d$, $1 \le j \le N$.

We acquire observations y_τ one by one, so that at time t we have at our disposal the observation $y^t = [y_1; ...; y_t]$. Our objective is to design tests $\mathcal{T}_t$ and thresholds $\rho_{tj} > 0$, $1 \le t \le d$, $1 \le j \le N$, meeting requirements completely similar to those from Section 3.2:

- given observation y^t, the test $\mathcal{T}_t$ should make either a nuisance, or a signal conclusion (but not both);
- if the input to (52) is a nuisance, the probability of the non-nuisance conclusion at time t should be at most ϵ_t, and if, for some j, the input is a signal of shape $j \le N$ and magnitude $> \rho_{tj}$, the probability of the nuisance conclusion at time t should be at most ϵ_{tj}.

Both these requirements should be satisfied for every $t \leq d$, and we would like to meet them with as small thresholds ρ_{tj} as possible.

We are about to demonstrate that the just outlined change detection problem can be handled via the techniques developed in Sections 3.3 and 3.4. Basically all we need to this end is to convert our observation scheme into the one considered in Section 3, and this is what we are about to do next.

4.2 Building the observation scheme

Given an input u, a noise ξ and $t \leq d$, the observation y^t is not uniquely determined by the input and the noise; it is also affected by the initial state x_0 of the system. To get rid of the influence of the initial condition we act as follows.

1. We denote by F^t the linear subspace of $E_y^t = \{[y_1; ...; y_t] \in \mathbf{R}^{tn_y}\}$ comprised of all outputs $y^t = [y_1; ...; y_t]$ which *in the noiseless case* $\xi = 0$ stem from zero input and some initial state of the system, build an orthonormal basis of the orthogonal complement of F^t in E_y^t and make the vectors of this basis the rows of a matrix, thus arriving at a $\mu_t \times (tn_y)$ matrix M_t. Note that μ_t is a nondecreasing function of t.

2. We set $z^t = M_t[y_1; ...; y_t]$, where y_τ is given by (52).

   It may happen that $\mu_t = 0$ for some t. In this case, our decision rule $\mathcal{T}_t$ by construction accepts the nuisance hypothesis, so that nontrivial decision rules will be associated only with those time instants t for which $\mu_t \geq 1$. Let $t = \kappa + 1$ be the first instant such that the corresponding $\mu_t \geq 1$. Note that time instants t with $\mu_t \geq 1$ form the final segment $\{\kappa + 1, \kappa + 2, ..., \kappa + K = d\}$ of $1, ..., d$ (recall that μ_t is nondecreasing in t). We set $m_k = \mu_{\kappa+k}$, $1 \leq k \leq K$. By construction,

$$z^{\kappa+k} = \overline{A}_k u + \overline{B}_k \xi, \tag{53}$$

   with some $m_k \times n_u$ matrix $\overline{A}_k$ and $m_k \times n_\xi$ matrix $\overline{B}_k$ readily given by our data.

From now on, we assume that[8] $K > 0$ and that $\overline{B}_k$ are of full row rank, i.e. $\mathrm{rank}(\overline{B}_k) = m_k$, $1 \leq k \leq K$. Finally, we select somehow invertible $m_k \times m_k$ matrices L_k and pass from observations (53) to observations

$$\omega^k = A_k u + \xi_k, \text{ where } \omega^k = L_k z^{\kappa+k}, \; A_k = L_k \overline{A}_k, \;\; \xi_k = Z_k \xi, \;\; Z_k = L_k \overline{B}_k. \tag{54}$$

For example, we can set $L_k = (\overline{B}_k \overline{B}_k^T)^{-1/2}$, thus ensuring that $Z_k Z_k^T = I_{m_k}$.

Notice that observations (54) meet the requirements imposed in Section 3.2. As a result, we find ourselves in the situation considered in Section 3.2 and therefore can apply to the change detection problem in question the machinery developed in Sections 3.3 and 3.4.

4.3 Illustration: detecting changes in the trend of a simple time series

We consider the example of Section 3.1 and apply, with minor modifications, the construction outlined in Section 3.4.

4.3.1 Constructing decision rules

To attain our goal (see Section 3.1) we act as follows. The rule $\mathcal{T}_1$ is trivial – it always accepts the nuisance hypothesis. To describe how $\mathcal{T}_k$, $k > 1$, is built, let us fix $k \in \{2, ..., d\}$.

1. *Building observation* ω^k. Setting $\xi = [\eta; \zeta]$ and $Z_k[\eta; \zeta] = \Theta_k^{-1}[D_k \eta + E_k \zeta]$, our observations (39) are as required in (40), and we meet Assumption A.

   In our implementation, we use $\Theta_k = \Omega_k^{1/2}$, where Ω_k is the minimum trace matrix satisfying $\Omega_k \succeq D_k D_k^T$, $\Omega_k \succeq E_k E_k^T$.

   We now apply the decision rules of Scheme II from Section 3.4 to observation (39). To define parameters $J_k, \rho_{kj}, \alpha_{1kj}$ and α_{2kj} we proceed as follows.

[8] Observe that when $K = 0$ our approach results in trivial tests always accepting the nuisance hypothesis – in this case the input-related component in the observations is fully masked by the influence of initial condition x_0.

2. We set
$$\widehat{J}_k = \{1, ..., 2k\};\ \widehat{\epsilon} = \tfrac{\epsilon}{d(d+1)-2};\ \epsilon_k = 2k\widehat{\epsilon};\ \epsilon_{kj} = \epsilon,\ j \in \widehat{J}_k;\ \rho_{kj} = R,\ j \notin \widehat{J}_k. \tag{55}$$
For $j \in \widehat{J}_k$, we specify α_{1kj}, α_{2kj}, δ_{kj} by the relations
$$\int_{\alpha_{1kj}}^{\infty} \gamma(s)ds = \widehat{\epsilon},\ \int_{\alpha_{2kj}}^{\infty} \gamma(s)ds = \epsilon_{kj} = \epsilon,\ \delta_{kj} = \tfrac{1}{2}[\alpha_{1kj} + \alpha_{2kj}].$$
Note that, by construction, setting $\epsilon_1 = 0$ we have
$$\forall k : \sum_{j\in\widehat{J}_k} \int_{\alpha_{1kj}}^{\infty} \gamma(s)ds \le \epsilon_k,\ \sum_{k=1}^{d} \epsilon_k = \epsilon \text{ and } \alpha_{1kj} + \alpha_{2kj} = 2\delta_{kj}, j \in \widehat{J}_k,$$
cf. (49).

3. For $j \in \widehat{J}_k$, we consider the convex optimization problem (cf. $(P_{kj}[\rho])$). $\mathrm{Opt}_{kj}(\rho)$ clearly is continuous and nonincreasing in $\rho > 0$ and $\lim_{\rho\to+0} \mathrm{Opt}(\rho) = 0$. When $\mathrm{Opt}_{kj}(R) \le \delta_{kj}$ we set $\rho_{kj} = R$. Otherwise, we find the smallest $\rho = \rho_{kj}$ such that $\mathrm{Opt}_{kj}(\rho) \ge \delta_{kj}$; note that in the latter case we have $\rho_{kj} \in (0, R)$ and $\mathrm{Opt}_{kj}(\rho_{kj}) = \delta_{kj}$. We have specified ρ_{kj} for all $j \in \widehat{J}_k$.

4. We set $J_k = \{j \in \widehat{J}_k : \rho_{kj} < R\}$ thus ensuring that $\rho_{kj} < R$ if and only if $j \in J_k$.[9]

5. Same as in Section 3.4, for $j \in J_k$ we denote by u_{kj} an optimal solution to problem $(P_{kj}[\rho])$ with $\rho = \rho_{kj}$ and set (cf. (50))
$$h_{kj} = -\frac{A_k u_{kj}}{\|A_k u_{kj}\|_2} = -\frac{A_k u_{kj}}{2\delta_{kj}},\ \ c_{kj} = \tfrac{1}{2}h_{kj}^T A_k u_{kj},\ \ \phi_{kj}(\omega) = h_{kj}^T\omega - c_{kj}.$$

6. Finally, given observation ω^k, our rule $\mathcal{T}_k$ makes the nuisance conclusion if and only if $\phi_{kj}(\omega_k) \ge \frac{1}{2}[\alpha_{2kj} - \alpha_{1kj}]$ for all $j \in J_k$, and makes signal conclusion otherwise, cf. Section 3.4. Invoking the results of Section 3.4.2, the decision rules we have built do satisfy risk specifications of Section 3.1.

4.3.2 Quantifying conservatism: performance indexes

Let us pass from intermediate observations (38) to observations
$$w^k = [D_k D_k^T]^{-1/2} z^k = \underbrace{Q_k\eta + S_k\zeta}_{\lambda^k} + Q_k u, \tag{56}$$
where $Q_k = [D_k D_k^T]^{-1/2} D_k$ satisfies $Q_k Q_k^T = I_{k-1}$ and $S_k = [D_k D_k^T]^{-1/2} E_k$. Since Q_k has orthonormal rows, specifying the distribution of η to be multivariate Student's $t_d(\nu, I_d)$ distribution on $\mathbf{R}^d$ with ν degrees of freedom and unit matrix parameter, the distribution $p(\cdot)$ of the random variable $Q_k\eta$ will be multivariate Student's $t_{k-1}(\nu, I_{k-1})$ distribution. Now, let θ be the smallest nonvanishing singular value of S_k (or, equivalently, θ^2 is the smallest eigenvalue of $S_k S_k^T$). Clearly, we can specify the covariance matrix Σ of zero mean d-dimensional Gaussian random vector ζ to satisfy $\Sigma \preceq \sigma^2 I_d$ and to be such that the covariance matrix of $S_k\zeta$ is $\theta^2\sigma^2 I_{k-1}$. We conclude that we can point out distributions of η and ζ, satisfying specifications of the model (36), and such that the random noise λ^k in (56) will be the sum of two independent zero-mean random vectors, one with $(k-1)$-dimensional Student distribution $t_{k-1}(\nu, I_{k-1})$, and the other – Gaussian, with covariance matrix $(\theta\sigma)^2 I_{k-1}$. As it is immediately seen, λ^k has a probability density $p(\cdot)$ of the form $f(\|\cdot\|_2)$ with nonincreasing f and whenever $e \in \mathbf{R}^{k-1}$ is a unit vector, the probability density $\gamma_k(\cdot)$ of the scalar random variable $e^T\lambda^k$ is
$$\gamma_k = \gamma_{\mathcal{S}} \star \gamma_{\theta\sigma}.$$

[9] Note that, typically, $|J_k| < |\widehat{J}_k|$. Thus, one can easily improve the estimation procedure by better accounting for the "remaining at step k" part of false alarm probability. A simple "dynamic" management of false alarm probabilities of tests is implemented in the numerical experiments described in the next section. The detailed construction is presented in the Online complement of the paper available at http://arxiv.org/abs/1705.07196.

Here, as above, $\gamma_{\mathcal{S}}$ is the density of the univariate Student's t_ν distribution, and $\gamma_{\theta_k\sigma}$ is the density of $\mathcal{N}(0,(\theta\sigma)^2)$.

Now let V_1, V_2 be two closed convex sets in the space $\mathbf{R}^d$ of inputs such that the sets Q_kV_χ, $\chi = 1, 2$, are closed, and one of these two sets is bounded, and let (u_*^1, u_*^2) be an optimal solution to the convex optimization problem

$$\delta = \min_{u^1\in V_1, u^2\in V_2} \tfrac{1}{2}\|Q_k[u^1 - u^2]\|_2. \tag{57}$$

By Remark 2.1, no test based on observation (56) can decide on two simple hypotheses $u = u_*^1$, $u = u_*^2$ with risk $< \int_\delta^\infty \gamma_k(s)ds$, implying that the same lower risk bound also holds true for all tests utilizing observations y^k rather than w^k.

Now let $V_1 = \{0\}$ and $V_2 = U_j(\rho)$, so that δ as defined in (57) becomes a (clearly, continuous and non-increasing when $\rho > 0$) function $\delta_{kj}(\rho)$ of ρ. Let us define ρ_{kj}^* as follows: if $\int_{\delta_{kj}(R)}^\infty \gamma_k(s)ds > \epsilon$, we set $\rho_{kj}^* = R$, otherwise $\rho_{kj}^* \in (0, R]$ is the smallest $\rho > 0$ such that $\int_{\delta_{kj}(\rho)}^\infty \gamma_k(s)ds \le \epsilon$. By construction, for every $\rho \in (0, \rho_{kj}^*)$ there is no test which, given an observation y^k, would decide with risk $\le \epsilon$ on the hypothesis "$u = 0$" vs. the alternative "u is a signal from $U_j(\rho)$ with $\rho > 0$." It is natural to quantify the conservatism of our decision rules $\mathcal{T}_k$ by the *performance indexes* ρ_{kj}/ρ_{kj}^*; the less are these indexes, the less is the conservatism.

4.3.3 Numerical results

We operate on time horizon $d = 8$ and deal with $\sigma = 1$ and with 5 values of the number ν of degrees of freedom of the Student distribution of η, specifically, the values $1, 2, 3, 6, \infty$. In the experiments of this section we use parameter values $R = 10^4$, $\epsilon = 0.01$.[10] The range of parameters ρ_{kj} and of ratios $\rho_{kj}/\rho_{k,i}^*$ are presented in Figure 1. Some comments are in order.

1. Relation $U_{2i-1}(\rho) = -U_{2i}(\rho)$ implies that $\rho_{k,2i-1} = \rho_{k,2i}$, $\rho_{k,2i-1}^* = \rho_{k,2i}^*$, $1 \le i \le d$.

2. We display the range of quantities ρ_{kj} and ρ_{kj}/ρ_{kj}^* only for those values of k, j for which $\rho_{kj}^* < R$, that is, ignore pairs k, j for which already an optimistic lower bounds ρ_{kj}^* on the magnitude of signal inputs of shape j which can be detected, with the required risk, at time k should be $\ge R$, which is forbidden by (37). On a closest inspection, the ignored pairs k, j are the pairs of the form $k, j = 2i - 1$ and $k, j = 2i$ where
   (a) $i > k$, or
   (b) $k = 1$, or
   (c) [only for pulses!] $i = 1 < k$.
   The reasons are clear: (a) stems from the fact that at time k it is impossible to detect a whatever large signal input which starts at time $i > k$. (b) reflects the fact that the contribution of a whatever large signal input, if any, to the very first observation is fully masked by the initial condition α_0, and in our model we do not impose any restrictions on this initial condition. (c) is of a similar origin: when the signal inputs are pulses, of a whatever magnitude, at time 1, are fully masked by the initial conditions and thus cannot be detected at all.

3. As it could be guessed, when the signal inputs are steps, the quantities ρ_{kj}, for j fixed, decrease as $k \ge j$ grows, since influence of step-change on our observations accumulates with time. In contrast, no such phenomenon is observed for pulse signal inputs where there is "nothing to accumulate."[11]

4. The conservatism of our decision rules, as presented in the tables, while unpleasant, seems to be not too high when $\nu \ge 3$, and becomes really arresting when $\nu = 1$. The origin of this phenomenon is quite transparent. The conservatism seems to stem primarily from systematic use in our constructions, for absence of something better, of the union bounds for probabilities. For example, when computing ρ_{kj}^*, we allow for the probability of false alarm at time k to be as large as ϵ, while in our decision rules, we "distribute" this probability between d instants where we make our decisions. Similarly, when computing ρ_{kj}^*, we act *as if* the only alternative to the nuisance hypothesis were a particular signal hypothesis $u \in U_j(\rho)$, while in fact we have several signal hypotheses to consider and should take into account the

[10] We use `Mosek` and YALMIP Matlab toolbox [1, 16] to solve corresponding optimization problems.

[11] Or, rather, that the noise in the states α_k of the model accumulates at the same rate, thus cancelling the effect of the growing observation sample.

resulting "accumulation of risk." As a result, we require from pairwise tests participating in $\mathcal{T}_k$ to have risk essentially smaller than ϵ, which in the case of a "heavy tail" noise distribution allowed by our model requires an essentially larger magnitudes of detectable signal inputs than those allowing for detection when the shape of signal input is known in advance. And indeed, we see that the ratios ρ_{kj}/ρ^*_{kj} rapidly increase as ν decreases.

We report on Figure 2 the evolution of parameters ρ_{kj} for $k = 8$ for step and pulse signals as a function of the risk of the test ϵ and the standard deviation σ of the Gaussian component ζ_t of the noise. As expected, ρ_{kj} increase with σ and when ϵ decreases.

4.3.4 Change detection in linear dynamic system revisited

The methodology developed in Section 3.4 allows for a straightforward refinement which hopefully improves the resulting inference performance. Note that the inference rules, as given in Sections 3.3 and 3.4 use very conservative bound for the probability of false alarm – for multiple tests this probability is simply the sum of probabilities of false rejections of the nuisance hypothesis for each test. This results in the increase of the testing thresholds ρ_{kj}, which amounts to a "logarithmic factor" in the case of Gaussian observation noise, but becomes much more severe in the case of a heavy-tail noise distribution. One way to make the decision less cautious is to reduce the number of hypotheses to test by aggregating the alternatives. Here we illustrate the idea of the proposed modification on the simple numerical example in section 4.3. Specifically, when building the detection procedure, at time k we act as follows:

- We compute the quantities ρ^*_{kj}, $j = 1, ..., 2d$, and denote by J^*_k the set of those j for which $\rho^*_{kj} < R$. As it was explained, there is no reason to bother to detect at time k signal inputs of shape $j \notin J^*_k$.
- Assume we have somehow associated thresholds $\rho_{kj} \geq \rho^*_{kj}$ to indexes $j \in J^*_k$; our goal, same as before, is to build a decision rule $\mathcal{T}_k$ which, given ω^k,
  — with probability at least $1-\epsilon$, makes signal conclusion at time k, provided the input belongs to $U_j(\rho_{kj})$ with some $j \in J^*_k$;
  — has false alarm probability at time k (the probability to make signal conclusion when the input is a nuisance) $\leq \epsilon_k$, $\sum_k \epsilon_k = \epsilon$.

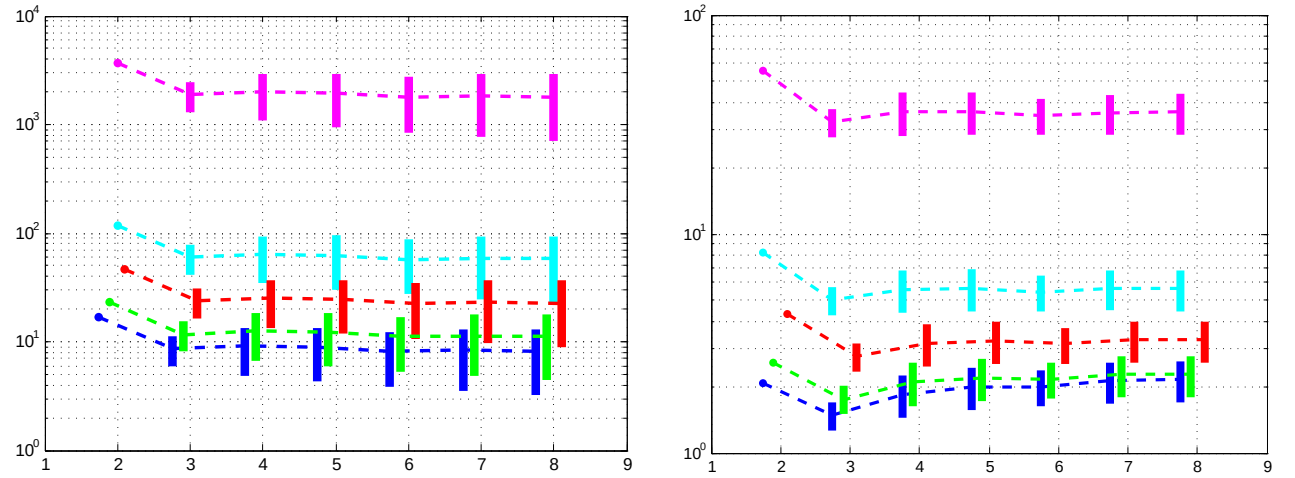
Next, let us color the sets $U_j(\rho_{kj})$, $j \in J^*_k$ (and their indexes) in a number L of colors; let I_ℓ be the set of indexes $j \in J^*_k$ colored by color ℓ. We associate with each ℓ a convex alternative U^ℓ – the convex hull of "alternatives of color ℓ":

$$U^\ell = \mathrm{Conv}\left(\bigcup_{j\in I_\ell} U_j(\rho_{kj})\right), \ \ell = 1, ..., L,$$

and replace the original detection problem with the following one: given observation ω^k we want to decide on the null hypothesis H_0 "the input is nuisance" (in our case, zero) vs. the alternative H_1 "the input belongs to $\bigcup_{\ell=1}^{L} U^\ell$." Same as before, our goal is to ensure probability of false alarm $\leq \epsilon_k$ and probability of miss $\leq \epsilon$.

Note that if we are able to do so, we meet our initial design specifications – with input from $U_j(\rho_{kj})$ for some $j \in J^*_k$, the probability of signal conclusion at time k will be at least $1-\epsilon$. To build the decision rule, let us use pairwise tests given by the construction from Section 3.4: we need L tests, ℓ-th of them deciding on H_0 vs. the alternative $H^\ell_1 : u \in U^\ell$, with risks ϵ^ℓ_k of false alarm and ϵ of miss. If we can ensure $\sum_{\ell=1}^{L} \epsilon^\ell_k \leq \epsilon_k$, we are done – the decision rule which makes nuisance conclusion when all our L tests "vote" for H_0, and makes the signal conclusion otherwise, is what we are looking for. Note that the original construction in Section 4.3 is of exactly this structure, with $L = \mathrm{Card}(J^*_k)$ (i.e., every $U_j(\rho_{kj})$, $j \in J^*_k$, has its own color). It could make sense, however, to use aggregated alternatives, thus reducing the number of colors. When doing so,

— on one hand, it is more difficult to decide on H_0 vs. the alternatives, because the image $A_k U^\ell$ of the "aggregated" set of signal inputs is closer to the image $\{0\}$ of the nuisance set than the images $A_k U_j(\rho_{kj})$, $j \in I_\ell$, of individual sets of signal inputs participating in the aggregation. Therefore, to ensure the same risk, we now need a somewhat larger separation of the images of the nuisance and the signal inputs in the observation space;



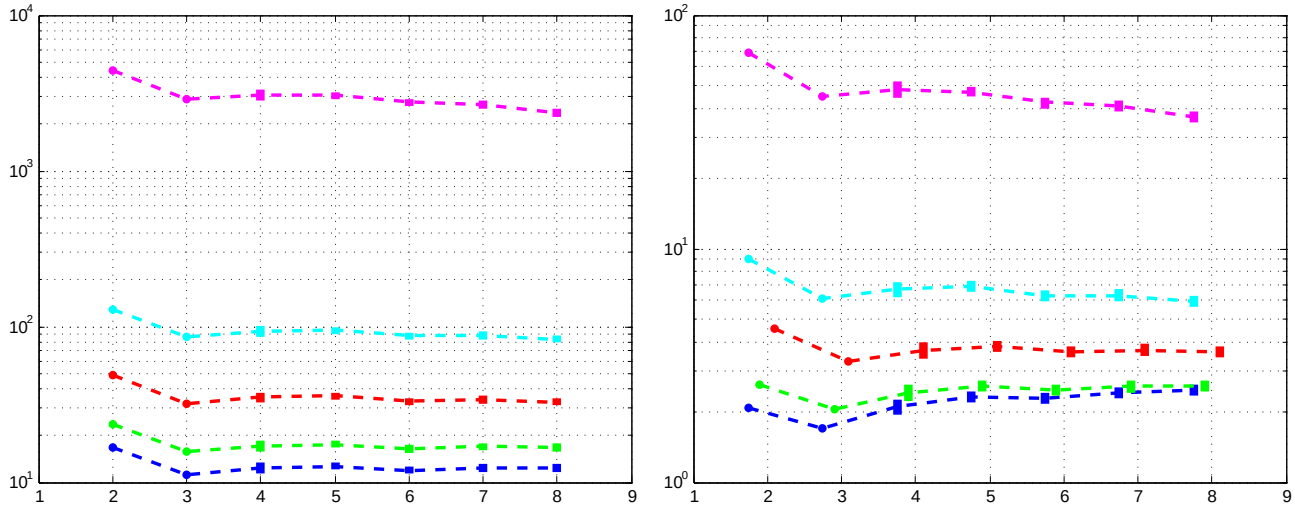
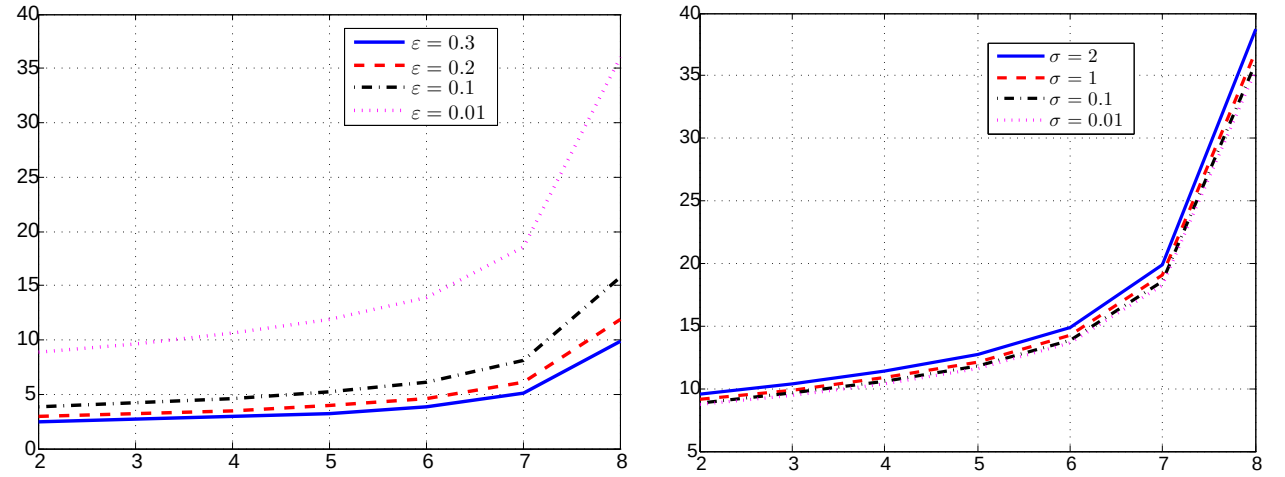
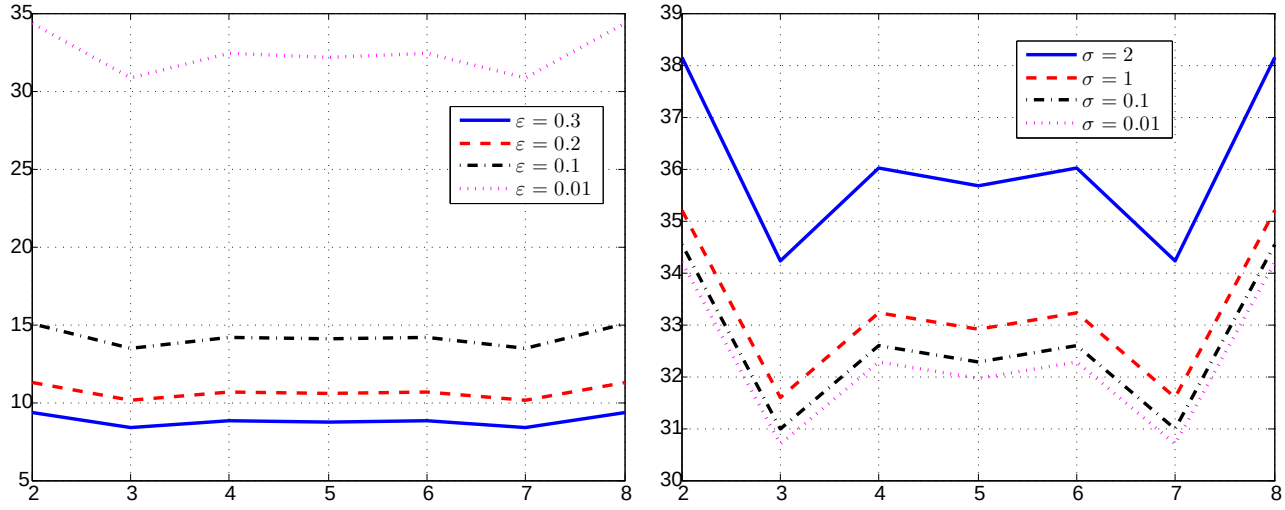


Figure 1: Detecting changes in the trend of a simple time series. Blue/green/red/cyan/magenta: $\nu = \infty/6/3/2/1$. Left plots: ranges (vertical segments) of $\rho_{k,2i-1} = \rho_{k,2i}$, vs. k. Right plots: ranges (vertical segments) of performance indexes $\rho_{k,2i-1}/\rho^*_{k,2i-1} = \rho_{k,2i}/\rho^*_{k,2i}$ vs. k. Ranges of i and k cover the domain where $\rho^*_{k,2i} = \rho^*_{k,2i-1} < R = 10^4$. Charts are shifted horizontally to improve the plot readability. On these plots both $\rho_{k,2i-1}$ and $\rho_{k,2i-1}/\rho^*_{k,2i-1}$ decrease with ν.



Signal input: step, $k = 8$, $2 \le i \le k$



Signal input: pulse, $k = 8$, $2 \le i \le k$

Figure 2: Coefficients $\rho_{k,2i-1} = \rho_{k,2i}$ for $\nu = 3$, $k = 8$, and $i = 2, \ldots, 8$, as a function of the desired risk ϵ of the test (left pane) and the standard deviation σ of the Gaussian noise ζ_t (right pane).

— on the other hand, we should now "distribute" ϵ_k among $L < \mathrm{Card}(J_k^*)$ miss probabilities ϵ_k^ℓ. This would allow to operate with larger miss probabilities, thus reducing the necessary separation of the images of the nuisance and the signal inputs in the observation space.

It is hard to tell in advance which of these two opposite effects will prevail; an answer, however, could be provided by computation, and it makes sense to give to the outlined modification a try. To make things as simple as possible, let us act as follows.

- After the colors are assigned and the sets U^ℓ, $\ell \le L$, are built, we specify ϵ_k^ℓ and the quantities α_{1k}, α_{2k}, δ_k to meet the requirements

$$\epsilon_k^\ell = \frac{\epsilon_k}{L},\ 1 \le \ell \le L;\ \int_{\alpha_{1k}}^{\infty} \gamma(s)ds = \frac{\epsilon_k}{L};\ \int_{\alpha_{2k}}^{\infty} \gamma(s)ds = \epsilon;\ \delta_k = \frac{1}{2}[\alpha_{1k} + \alpha_{2k}].$$

Let us suppose that the relation

$$\min_{u \in U^\ell} \tfrac{1}{2}\|A_k u\|_2 \ge \delta_k,\ \ell = 1, ..., L, \tag{58}$$

is satisfied. We set

$$h_{k\ell} = A_k u_{k\ell}/\|A_k u_{k\ell}\|_2,\ \ c_{k\ell} = \tfrac{1}{2} h_{k\ell}^T A_k u_{k\ell},\ \ \phi_{k\ell}(\omega^k) = h_{k\ell}^T \omega_k - c_{k\ell},$$

where $u_{k\ell}$ are optimal solutions to the optimization problems in (58). It is immediately seen that making at time k the nuisance conclusion if and only if $\phi_{kj}(\omega^k) \ge \frac{1}{2}[\alpha_{2k} - \alpha_{1k}]$, we ensure simultaneously the probability of false alarm at time k at most ϵ_k, and the probability of miss when the input belongs to $\bigcup_{j \in J_k^*} U_j(\rho_{kj})$ at most ϵ, thus meeting our design specifications.

Specifying ρ_{kj}. The question we did not address so far is how to choose ρ_{kj}, $j \in J_k^*$. What we expect of these quantities is to ensure the validity of (58), and the simplest way to achieve this goal is as follows. Setting $\rho_{kj} = \theta \rho_{kj}^*$, the left hand side in (58) is a nondecreasing function of θ, and we can find by bisection the smallest $\theta = \theta_k \ge 1$ for which (58) takes place. After θ_k is found, we set $\rho_{kj} = \theta_k \rho_{kj}^*$, $j \in J_k^*$. Clearly, with this approach, the performance indexes ρ_{kj}/ρ_{kj}^*, $j \in J_k^*$, are all equal to θ_k.

How it works. We applied the just outlined construction to the data underlying the numerical experiment described in Section 4.3. Our implementation was the simplest possible: given k, we looked at all j's such that $\rho_{kj}^* < R$; the set J_k^* of these j's with our data is nonempty only when $k \ge 2$ and is either $\{1, ..., 2k\}$ (step signals), of $\{3, 4, ..., 2k\}$ (pulse signals). A set $U_j(\rho)$ with odd index $j = 2i - 1$ (even index $j = 2i$), $j \in J_k^*$, is comprised of signals which are zero before time i and "jump up/jump down" at time i depending on whether j is even or odd. We color these sets in $L = 2$ colors, depending on whether the corresponding indexes j are odd or even, that is, we use at step k

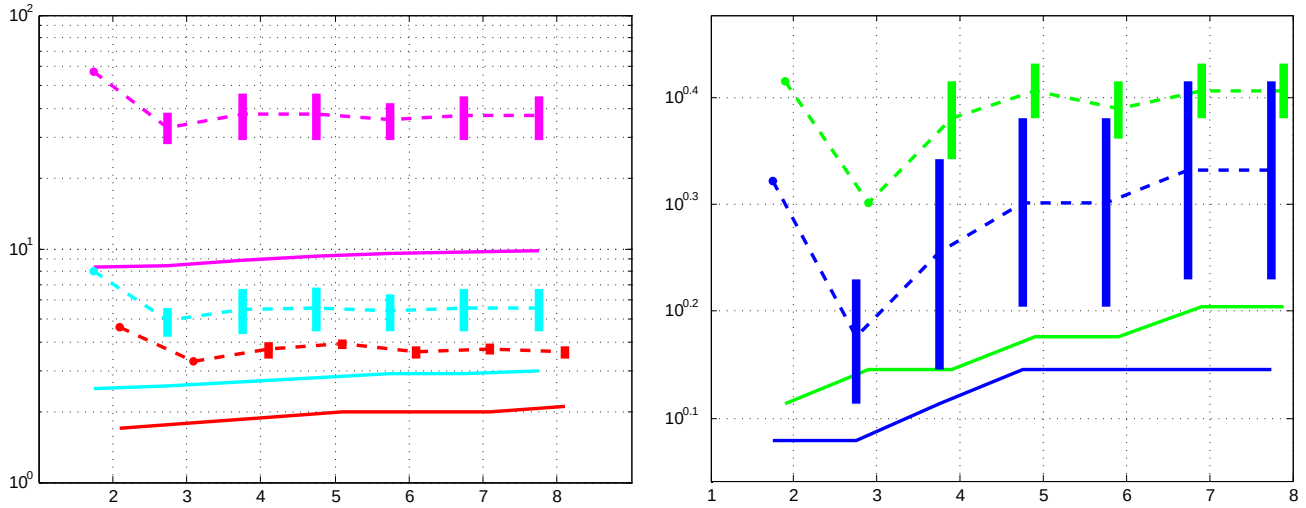
$$U^\ell = \bigcup_{j \in J_k^*,\ j \bmod 2 = \ell} U_j(\theta_k \rho_{kj}^*),\ \ell = 1, 2,$$

with θ_k as explained above.

We present on Figure 3 the comparison of performance indexes ρ_{kj}/ρ_{kj}^*, $j \in J_k^*$, for the original inference routine (these indexes are presented on Figure 1) and the performance indexes of the just described modified inference. For our initial routine, the performance indexes slightly vary with $j \in J_k^*$, and we present their ranges; for the new routine, the performance indexes do not depend on j. We observe that in the considered example the proposed straightforward aggregation of signal inputs typically results in degradation of the performance indexes, and improves these indexes significantly for the "heavy tailed" noise distributions (the case of $\nu = 1$).

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/.



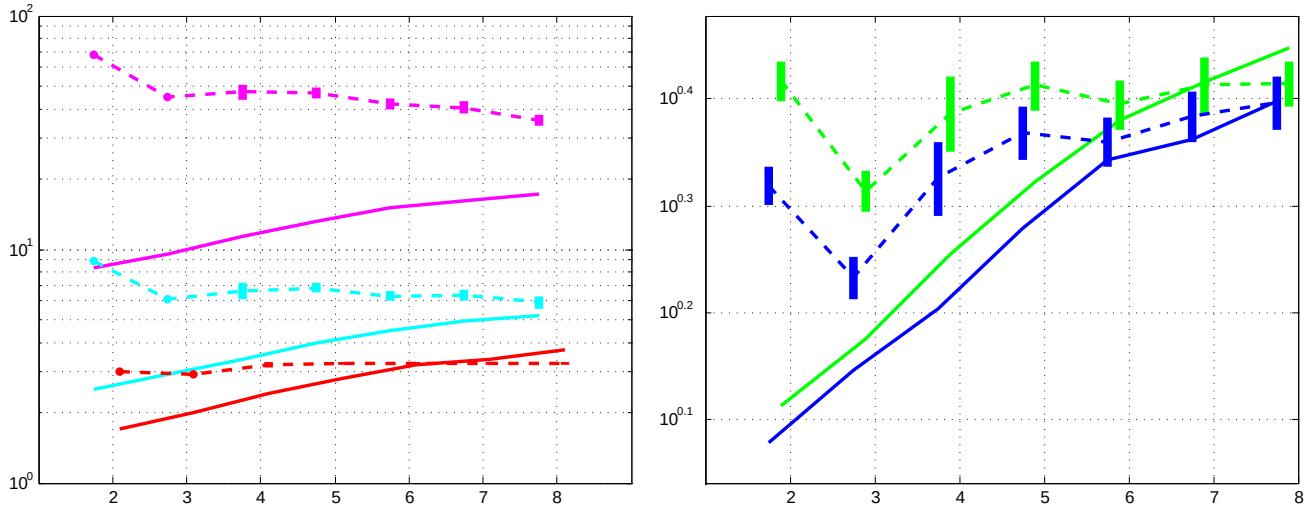


Figure 3: Ranges of performance indexes ρ_{kj}/ρ^*_{kj} (cf. right plots of Figure 1) as compared to the performance index for the refined inference (solid lines). Red/cyan/magenta: $\nu = 3/2/1$ (left plots), blue/green: $\nu = \infty/6$ (right plots). Charts are shifted horizontally to improve the plot readability. On these plots $\rho_{k,j}/\rho^*_{k,j}$ decrease with ν.

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

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

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

[5] Y. Cao, V. Guigues, A. Juditsky, A. Nemirovski, and Y. Xie. Change detection via affine and quadratic detectors. *arXiv preprint arXiv:1608.00524*, 2016.

[6] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952.

[7] L. Dumbgen and V. G. Spokoiny. Multiscale testing of qualitative hypotheses. *Annals of Statistics*, pages 124–152, 2001.

[8] T. Eltoft, T. Kim, and T.-W. Lee. On the multivariate laplace distribution. *IEEE Signal Processing Letters*, 13(5):300–303, 2006.

[9] A. Goldenshluger, A. Juditsky, and A. Nemirovski. Hypothesis testing by convex optimization. *Electronic Journal of Statistics*, 9(2):1645–1712, 2015.

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

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

[12] A. Juditski and A. Nemirovski. On sequential hypotheses testing via convex optimization. *Automation and Remote Control*, 76:809–825, 2015.

[13] A. Juditsky and A. Nemirovski. Hypothesis testing via affine detectors. *Electronic Journal of Statistics*, 10:2204–2242, 2016.

[14] S. Kotz and S. Nadarajah. *Multivariate t-distributions and their applications*. Cambridge University Press, 2004.

[15] E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[16] J. Löfberg. Yalmip : A toolbox for modeling and optimization in matlab. *In Proceedings of the IEEE CACSD Conference*, 2004.

A Proofs

A.1 Proof of Proposition 2.3

In what follows, for functions $f, g : \mathbf{R} \to \mathbf{R}$, we say that f dominates g (notation: $f \succeq g$, or, equivalently, $g \preceq f$), if

$$\int_\delta^\infty f(s)ds \geq \int_\delta^\infty g(s)ds, \ \forall \delta \geq 0.$$

Let

- $\mathcal{E}$ be the family of even probability densities on the real axis,
- $\mathcal{N}$ be the family of nice functions on the axis.

1^o. Note that $\succeq$ clearly is transitive: if $f \succeq g$ and $g \succeq h$, then $f \succeq h$. Furthermore, $\mathcal{N} \subset \mathcal{E}$ (by definition of $\mathcal{N}$), and $\mathcal{E}$ is closed with respect to taking convolution (evident). We also need the following technical facts.

1^o.a $\mathcal{N}$ is closed with respect to taking convolution.

Indeed, let $f, g \in \mathcal{N}$. The fact that $f \star g$ is even and continuous on the real axis is evident. Therefore in order to show that $f \star g \in \mathcal{N}$ it suffices to verify that $h = f \star g$ is nonincreasing on the nonnegative ray. For $0 < z \le f(0)$, denoting $\bar{y}(z) = \min\{y \ge 0 : f(y) = z\}$, we have for every $x \ge 0$,

$$h(x) = \int_{-\infty}^{\infty} f(x-y)g(y)dy = \int_{-\infty}^{\infty} \Big(\int_0^{f(x-y)} dz\Big) g(y)dy = \int_0^{f(0)} H(x,z)dz$$

where $H(x,z) = \int_{x-\bar{y}(z)}^{x+\bar{y}(z)} g(y)dy$ for $0 < z \le f(0)$ and $H(x,0) = 0$. To conclude we observe that for every fixed $f(0) \ge z \ge 0$, $H(\cdot, z)$ is a differentiable, nonnegative, and nonincreasing function on the nonnegative ray. Indeed, for $0 < z \le f(0)$, the derivative $H'_x(x,z)$ of this function at x is $H'_x(x,z) = g(x+\bar{y}(z)) - g(x-\bar{y}(z))$ and it is clear that this quantity is ≤ 0 since $g \in \mathcal{N}$. We have checked that for every fixed $z \ge 0$, $H(\cdot, z)$ is nonincreasing on the nonnegative ray. It follows that h is nonincreasing on the nonnegative ray.

1^o.b Let $\bar{f}, f \in \mathcal{E}$, $g \in \mathcal{N}$, and let $\bar{f} \succeq f$. Then $\bar{f} \star g \succeq f \star g$.

Let us verify that for all $x \ge 0$, $\bar{H}(x) \le H(x)$, where $\bar{H}$ and H are cumulative distribution functions (c.d.f.) of the densities $\bar{f} \star g$ and $f \star g$, respectively. Observe that

$$H(x) = \int_{-\infty}^{\infty} g(s)F(x-s)ds = \int_{-\infty}^{\infty} g(x-t)F(t)dt,$$

same as

$$\bar{H}(x) = \int_{-\infty}^{\infty} g(x-t)\bar{F}(t)dt,$$

where F and $\bar{F}$ are the c.d.f.'s of the densities f and $\bar{f}$, respectively. Thus, when setting $\Delta(s) = F(s) - \bar{F}(s)$, we get

$$\begin{array}{rcl} H(x) - \bar{H}(x) & = & \int_{-\infty}^{\infty} g(x-t)\Delta(t)dt = \int_0^{\infty} g(x-t)\Delta(t)dt + \int_0^{\infty} g(x+t)\Delta(-t)dt \\ [\text{because } \Delta(-t) = -\Delta(t)] & = & \int_0^{\infty} \Delta(t)[g(x-t) - g(x+t)]dt \ge 0, \end{array}$$

where the final "$\ge$" is due to $\Delta(t) \ge 0$ and $g(x-t) \ge g(x+t)$ for $x, t \ge 0$ (recall that $g \in \mathcal{N}$) for $t \ge 0$.

1^o.c. Let $f \in \mathcal{E}$, $\rho \in (0,1)$, and let $f_\rho(s) = \rho^{-1} f(\rho^{-1}s)$. Then $f_\rho \in \mathcal{E}$ and $f_\rho \preceq f$.

Indeed, the inclusion $f_\rho \in \mathcal{E}$ is obvious. On the other hand, if $\xi \sim f$, one has $P(\rho\xi \ge x) = P(\xi \ge x/\rho) \le P(\xi \ge x)$ for $0 < \rho < 1$ and $x \ge 0$, what is exactly $f_\rho \preceq f$.

1^o.d. Let $f \in \mathcal{N}$ and $g \in \mathcal{E}$. Then $f \star g \succeq f$.

Note that the c.d.f. H of $f \star g$ satisfies

$$H(x) = \int_{-\infty}^{\infty} g(s)F(x-s)ds = \int_0^{\infty} g(s)[F(x-s) + F(x+s)]ds,$$

where F is the c.d.f. of f (recall that $g \in \mathcal{E}$), and therefore for $x \ge 0$ it holds

$$H(x) - F(x) = \int_0^{\infty} g(s)[F(x-s) + F(x+s) - 2F(x)]ds \le 0$$

due to the concavity of F on $\mathbf{R}_+$.

2^0. Now let $p(\cdot) \in \mathcal{P}^n$, $q(\cdot) \in \mathcal{P}^m$, and let $e \in \mathbf{R}^r$, $\|e\|_2 = 1$, be given. Let us set $e_1 = A^T\Theta^{-1}e$, $e_2 = B^T\Theta^{-1}e$, let $\eta \sim p$ and $\zeta \sim q$ be independent, and let $\xi = \Theta^{-1}[A\eta + B\zeta]$. We have

$$\omega := e^T\xi = [A^T\Theta^{-1}e]^T\eta + [B^T\Theta^{-1}e]^T\zeta = e_1^T\eta + e_2^T\zeta.$$

Observe that $e_1^Te_1 = e^T\Theta^{-1}AA^T\Theta^{-1}e \le 1$ due to $\Theta^{-1}AA^T\Theta^{-1} \preceq I_r$, and for similar reasons $e_2^Te_2 \le 1$. We denote $\rho_1 = \|e_1\|_2$ and $\rho_2 = \|e_2\|_2$, so that $\rho_1, \rho_2 \in [0,1]$. When $\rho_\chi > 0$, $\chi = 1, 2$, we set $\bar{e}_\chi = \rho_\chi^{-1}e_\chi$. Now, let $f_1 \in \mathcal{E}$ (respectively, $f_2 \in \mathcal{E}$) be the density of the scalar random variable $e_1^T\eta$ (respectively, $e_2^T\zeta$). Note that when $\rho_1 > 0$ ($\rho_2 > 0$) random variable $e_1^T\eta$ ($e_2^T\zeta$) indeed has a density, and this density is even. Let also $\bar{f}_1 \in \mathcal{E}$ ($\bar{f}_2 \in \mathcal{E}$) be the density of $\bar{e}_1^T\eta$ ($\bar{e}_2^T\zeta$).

2^0.a. Assume for a moment that $\rho_1 > 0$ and $\rho_2 > 0$. Then

- $p \in \mathcal{P}_\mu^n$, whence $f_1 \in \mathcal{E}$, $\bar{f}_1 \in \mathcal{E}$ and $f_1(s) = \rho_1^{-1}\bar{f}_1(\rho_1^{-1}s)$, whence $f_1 \preceq \bar{f}_1$ by $\mathbf{1}^o$.**c**. Since $\bar{f}_1 \preceq \mu$ and $\preceq$ is transitive, we have also $f_1 \preceq \mu$.
- $q \in \mathcal{P}^m$, and $\mathcal{P}^m$ is a completely monotone subfamily of $\mathcal{P}_\nu^m$, whence $f_2 \in \mathcal{N}$, $\bar{f}_2 \in \mathcal{N}$ and, same as above, $f_2 \preceq \bar{f}_2 \preceq \nu$, whence also $f_2 \preceq \nu$.
- The density of ω is $f_1 \star f_2$. We have

$$\begin{array}{ll} f_1 \star f_2 \preceq \mu \star f_2 & [\text{by } \mathbf{1}^o.\mathbf{b} \text{ in view of } f_1, \mu \in \mathcal{E},\ f_2 \in \mathcal{N} \text{ and } f_1 \preceq \mu], \\ f_2 \star \mu \preceq \mu \star \nu & [\text{by } \mathbf{1}^o.\mathbf{b} \text{ in view of } f_2, \nu \in \mathcal{E},\ \mu \in \mathcal{N} \text{ and } f_2 \preceq \nu]. \end{array}$$

  Hence, $f_1 \star f_2 \preceq \gamma := \mu \star \nu$, such that $\gamma \in \mathcal{E}$, and (15) follows. Besides this, in the case in question ω has an even density.

2^0.b. Now let $\rho_1 = 0$. Then $\rho_2 > 0$ due to $AA^T + BB^T \succ 0$, and the probability density of ω is f_2. We have $f_2 \preceq \bar{f}_2 \preceq \nu \preceq \mu \star \nu$ (the concluding $\preceq$ is due to $\mathbf{1}^o$.**d**), and (15) follows. Similarly, when $\rho_2 = 0$, we have $\rho_1 > 0$, and the probability density of ω is f_1. Similarly to the case where $\rho_1 = 0$, we have $f_1 \preceq \bar{f}_1 \preceq \mu \preceq \mu \star \nu$, and (15) follows. Furthermore, as we have seen, ω always has an even probability density. Finally, $\gamma = \mu \star \nu$ is nice by $\mathbf{1}^o$.**a**. (i) is proved.

3^0. To prove (ii), note that in the notation of item 2^0 and under the premise of (ii), $\rho_\chi > 0$ implies that $f_\chi \in \mathcal{N}$. Hence, due to $\mathbf{1}^o$.**a**, the distribution of ω has density from $\mathcal{N}$ (this density is $f_1 \star f_2$ when $\rho_1, \rho_2 > 0$, f_2 when $\rho_1 = 0$, and f_1 when $\rho_2 = 0$), which combines with (i) to imply (ii). □

A.2 Proof of Proposition 2.4

Proof. When $x \in X_1$, we have $h_*^T x \geq h_*^T x_*^1 = c_* + \delta$, thus

$$\begin{array}{rcl} \{\xi : s_*(x+\xi) < \frac{1}{2}(\alpha_2 - \alpha_1)\} & = & \{\xi : h_*^T(x+\xi) < c_* + \frac{1}{2}(\alpha_2 - \alpha_1)\} \\ & = & \{\xi : h_*^T\xi < c_* - h_*^T x + \frac{1}{2}(\alpha_2 - \alpha_1)\} \\ & \subseteq & \{\xi : h_*^T\xi < c_* - h_*^T x_*^1 + \frac{1}{2}(\alpha_2 - \alpha_1)\} \\ & = & \{\xi : h_*^T\xi < -\delta + \frac{1}{2}(\alpha_2 - \alpha_1)\} \subseteq \{h_*^T\xi < -\alpha_1\}, \end{array}$$

where the last inclusion is due to $\alpha_1 + \alpha_2 \leq 2\delta$. Therefore, for $p \in \mathcal{P}$ it holds

$$\mathrm{Prob}_{\xi\sim p}\{s_*(x+\xi) < \tfrac{1}{2}(\alpha_2 - \alpha_1)\} \leq \int\limits_{h_*^T\xi < -\alpha_1} p(\xi)d\xi \underbrace{=}_{(a)} \int\limits_{h_*^T\xi > \alpha_1} p(\xi)d\xi \underbrace{\leq}_{(b)} P_\gamma(\alpha_1)$$

where (a) is due to the fact that $p(\cdot)$ is even, and (b) is due to $p \in \mathcal{P}_\gamma$ and $\alpha_1 \geq 0$. When $x \in X_2$, we have $h_*^T x \leq h_*^T x_*^2 = c_* - \delta$, and using a completely similar argument we conclude that for $p \in \mathcal{P}$ it holds

$$\mathrm{Prob}_{\xi\sim p}\{s_*(x+\xi) \geq \tfrac{1}{2}(\alpha_2 - \alpha_1)\} \leq P_\gamma(\alpha_2).$$ □

A.3 Proof of Proposition 2.5

When $x_k \in X_1^k$ and $p_k \in \mathcal{P}^k$ for all $k \leq K$, due to the origin of h_k, δ_k and c_k, we have

$$h_k^T(x_k + \xi_k) - c_k \geq \delta_k + h_k^T\xi_k,$$

and, because $\eta_k(\cdot)$ is nondecreasing,

$$\int \mathrm{e}^{-\eta_k(h_k^T[x_k+\xi_k]-c_k)}p_k(\xi_k)d\xi_k \leq \int \mathrm{e}^{-\eta_k(\delta_k+h_k^T\xi_k)}p_k(\xi_k)d\xi_k \leq \mathrm{risk}_{\delta_k}(\eta_k|\mathcal{P}^k), \quad k = 1, ..., K,$$

where the concluding "$\le$" is due to (23.a). Hence,

$$\begin{aligned}\int \mathrm{e}^{-\phi^{(K)}(x_1+\xi_1,\ldots,x_K+\xi_K)} \textstyle\prod_{k=1}^K [p_k(\xi_k)d\xi_k] &= \prod_{k=1}^K \left[\int \mathrm{e}^{-\phi_k(x_k+\xi_k)} p_k(\xi_k)d\xi_k\right] \\ &= \prod_{k=1}^K \left[\int \mathrm{e}^{-\eta_k(h_k^T[x_k+\xi_k]-c_k)} p_k(\xi_k)d\xi_k\right] \le \textstyle\prod_{k=1}^K \left[\mathrm{risk}_{\delta_k}(\eta_k|\mathcal{P}^k)\right].\end{aligned}$$

When $x_k \in X_2^k$ and $p_k \in \mathcal{P}^k$ for all $k \le K$, in a completely similar way we obtain

$$\int \mathrm{e}^{\phi^{(K)}(x_1+\xi_1,\ldots,x_K+\xi_K)} \prod_{k=1}^K [p_k(\xi_k)d\xi_k] \le \prod_{k=1}^K \left[\mathrm{risk}_{\delta_k}(\eta_k|\mathcal{P}^k)\right]. \qquad \square$$

A.4 Proof of Proposition 2.6

Let $p(\cdot) \in \mathcal{P}$, let $e \in \mathbf{R}^n$ be a unit vector, and let $q(\cdot)$ be the probability density of the scalar random variable $e^T\xi$ induced by the density $p(\cdot)$ of ξ. We start with the following well known observation:

Lemma A.1 *Let f, g be two probability densities on $\mathbf{R}$ such that*

$$\begin{array}{ll}(a) & \int_0^\infty f(s)ds = \int_0^\infty g(s)ds, \\ (b) & \int_r^\infty f(s)ds \ge \int_r^\infty g(s)ds, \quad \forall r \ge 0,\end{array} \tag{59}$$

and let $h(s)$ be a nondecreasing real-valued function on the nonnegative ray such that $\int_0^\infty h(s)f(s)ds < \infty$. Then

$$\int_0^\infty h(s)f(s)ds \ge \int_0^\infty h(s)g(s)ds. \tag{60}$$

To make the presentation self-contained, here is the proof of the lemma:

In view of (59.a), we can assume w.l.o.g. that $h(0) = 0$. Let us extend $h(s)$ from the nonnegative ray to the entire real axis by setting $h(s) = 0$, $s < 0$, thus arriving at a monotone on the axis nonnegative function. Let $\eta \sim f$ and $\zeta \sim g$. When denoting H the c.d.f. of $h(\eta)$, under the premise of the lemma we clearly have

$$\mathbf{E}_{\eta\sim f}\{h(\eta)\} = \int_0^\infty t dH(t) = \int_0^\infty (1 - H(t))dt = \int_0^\infty \mathrm{Prob}\{h(\eta) > t\}dt.$$

On the other hand, for $t \ge 0$ the set $\{s : \ h(s) > t\}$ is a ray either of the form $[a_t, +\infty)$ or $(a_t, +\infty)$ with $a_t \ge 0$, so that (59.b) implies that

$$\forall t \ge 0 \ \ \mathrm{Prob}\{h(\eta) > t\} \ge \mathrm{Prob}\{h(\zeta) > t\}.$$

As a result,

$$\mathbf{E}_{\eta\sim f}\{h(\eta)\} = \int_0^\infty \mathrm{Prob}\{h(\eta) > t\}dt \ge \int_0^\infty \mathrm{Prob}\{h(\zeta) > t\}dt.$$

We conclude that $\mathbf{E}_{\zeta\sim g}\{h(\zeta)\}$ is finite and satisfies

$$\mathbf{E}_{\zeta\sim g}\{h(\zeta)\} = \int_0^\infty \mathrm{Prob}\{h(\zeta) > t\}dt \le \mathbf{E}_{\eta\sim f}\{h(\eta)\}. \qquad \square$$

1^0. Note that $q(\cdot)$ is an even probability density on the axis and

$$\int_s^\infty [\gamma(r)-q(r)]dr = P_\gamma(s) - \mathrm{Prob}_{\xi\sim p}\{\xi:\ e^T\xi \geq s\} \left\{ \begin{array}{ll} \geq 0, & s\geq 0,\\ = 0, & s = 0.\end{array}\right.$$

We have

$$\int_{\mathbf{R}^n} \mathrm{e}^{-\eta(\delta+e^T\xi)}p(\xi)d\xi = \int_{-\infty}^\infty \mathrm{e}^{-\eta(\delta+s)}q(s)ds \underbrace{=}_{(a)} \int_0^\infty H_{\delta\eta}(s)q(s)ds \underbrace{\leq}_{(b)} \int_0^\infty H_{\delta\eta}(s)\gamma(s)ds \underbrace{=}_{(c)} \epsilon_\delta(\eta|\gamma),$$

where (a) is due to the fact that q is even, (b) is a result of applying Lemma A.1 to densities γ, q and nondecreasing $H_{\delta\eta}$, and (c) is due to the definition (26) of the δ-index.

2^0. We have

$$\begin{array}{rl} & \int_{\mathbf{R}^n} \mathrm{e}^{\eta(-\delta+e^T\xi)}p(\xi)d\xi = \int_{\mathbf{R}^n} \mathrm{e}^{-\eta(\delta-e^T\xi)}p(\xi)d\xi \ [\text{since } \eta(\cdot) \text{ is odd}]\\ = & \int_{\mathbf{R}^n} \mathrm{e}^{-\eta(\delta+e^T\xi)}p(\xi)d\xi \ [\text{since } p(\cdot) \text{ is even}],\end{array}$$

and we have already seen in $\mathbf{1}^0$ that the concluding quantity is $\leq \epsilon_\delta(\eta|\gamma)$. The bottom line is that inequalities (23) hold true with $\epsilon = \epsilon_\delta(\eta|\gamma)$, and (27) follows. □

A.5 Proof of Proposition 2.7

Let $e \in \mathbf{R}^n$ be a unit vector, $\delta \geq 0$, and $p \in \mathcal{P}_{\mathrm{sG}}$. We have

$$\int_{\mathbf{R}^n} \mathrm{e}^{-\eta(\delta+e^T\xi)}p(\xi)d\xi = \int_{\mathbf{R}^n} \mathrm{e}^{-\delta^2-\delta e^T\xi}p(\xi)d\xi \leq \mathrm{e}^{-\delta^2+\delta^2/2} = \mathrm{e}^{-\delta^2/2},$$

where the concluding $\leq$ is due to $p \in \mathcal{P}_{\mathrm{sG}}$ and $\|e\|_2 = 1$. Similarly,

$$\int_{\mathbf{R}^n} \mathrm{e}^{\eta(e^T\xi-\delta)}p(\xi)d\xi = \int_{\mathbf{R}^n} \mathrm{e}^{-\delta^2+\delta e^T\xi}p(\xi)d\xi \leq \mathrm{e}^{-\delta^2+\delta^2/2} = \mathrm{e}^{-\delta^2/2}.$$

The resulting inequalities hold true for all unit vectors e and all $p \in \mathcal{P}_{\mathrm{sG}}$, implying (33). □

A.6 Proof of Proposition 2.8

1^o. Let $p \in \mathcal{P}$ and $x \in X_1$ be fixed. Due to the monotonicity of η, and by the definition of c_*, we get from (34.a):

$$\mathbf{E}_{\xi\sim p}\{\eta(h_*^T(x+\xi)+c)\} \geq \mathbf{E}_{\xi\sim p}\{\eta(h_*^T(x_*^1+\xi)+c)\} = c_* + \frac{\mathrm{e}^\kappa\varrho}{1+\mathrm{e}^\kappa},$$

and so

$$\mathbf{E}_{\xi^K\sim p\times...\times p}\psi_j([x+\xi_1;...;x+\xi_K]) = \mathbf{E}_{\xi\sim p}\{\eta(h^T(x+\xi)+c)\} \geq c_* + \frac{\mathrm{e}^\kappa\varrho}{1+\mathrm{e}^\kappa}.$$

On the other hand, by (34.b) we have

$$\mathbf{Var}_{\xi^K\sim p\times...\times p}(\psi_j(\omega^K)) = m^{-1}\mathbf{Var}_{\xi\sim p}\{\eta(h^T(x+\xi)+c)\} \leq m^{-1}.$$

Now, by the Chebyshev inequality,

$$\begin{array}{rcl}\mathrm{Prob}_{\xi^K\sim p\times...\times p}\left\{\psi_j(\omega^K) < c_*\right\} & \leq & \mathrm{Prob}_{\xi^K\sim p\times...\times p}\left\{\psi_j(\omega^K) - \mathbf{E}_{\xi^K\sim p\times...\times p}\{\psi_j(\omega^K)\} < -\frac{\mathrm{e}^\kappa\varrho}{1+\mathrm{e}^\kappa}\right\}\\ & \leq & \frac{(1+\mathrm{e}^\kappa)^2}{m\mathrm{e}^{2\kappa}\varrho^2} \leq \frac{1}{4\mathrm{e}}.\end{array}$$

As a result, the risk $\mathrm{Risk}_{1\mathcal{S}}$ of the test $\mathcal{T}_K^{\rm mm}$ satisfies the bound

$$\mathrm{Risk}_{1\mathcal{S}}(\mathcal{T}_K^{\rm mm}|\mathcal{P},X_1,X_2)\le\sum_{J/2\le j\le J}\left(\begin{array}{c}J\\j\end{array}\right)(4\mathrm{e})^{-j}\left(1-\frac{1}{4\mathrm{e}}\right)^{J-j}\le 2^J(4\mathrm{e})^{-J/2}=e^{-J/2}\le\epsilon_1,$$

where the final inequality is due to (35).
2^o. The same argument, as applied to $x\in X_2$ results in

$$\mathbf{E}_{\xi^K\sim p\times...\times p}\psi_j([x+\xi_1;...;x+\xi_K])=\mathbf{E}_{\xi\sim p}\{\eta(h_*^T(x+\xi)+c)\}\le c_*-\frac{\varrho}{1+\mathrm{e}^\kappa},$$

and

$$\begin{array}{rcl}\mathrm{Prob}_{\xi^K\sim p\times...\times p}\left\{\psi_j(\omega^K)\ge c_*\right\}&\le&\mathrm{Prob}_{\xi^K\sim p\times...\times p}\left\{\psi_j(\omega^K)-\mathbf{E}_{\xi^K\sim p\times...\times p}\{\psi_j(\omega^K)\}\ge\frac{\varrho}{1+\mathrm{e}^\kappa}\right\}\\&\le&\frac{(1+\mathrm{e}^\kappa)^2}{m\varrho^2}\le\frac{1}{4}\mathrm{e}^{2\kappa-1}.\end{array}$$

Same as above, we conclude that

$$\mathrm{Risk}_{2\mathcal{S}}(\mathcal{T}_K^{\rm mm}|\mathcal{P},X_1,X_2)\le\mathrm{e}^{J\left(\kappa-\frac{1}{2}\right)}=\exp\left(-\frac{J}{2}\frac{\ln\epsilon_2^{-1}}{\ln\epsilon_1^{-1}}\right)\le\epsilon_2,$$

where the concluding inequality is due to (35). □