Hypothesis Testing via Convex Optimization

Arkadi Nemirovski
Joint research with
Alexander Goldenshluger
and
Anatoli Iouditski

IPPI, Moscow, December 2014

♠ Consider two results of High-Dimensional Statistics:

Theorem A [Ibragimov & Khas'minskii 1979] *Given α, L, k, let $\mathcal{X}$ be the set of all functions $f : [0,1] \to \mathbb{R}$ with (α, L)-Hölder continuous k-th derivative. The minimax risk of recovering $f(0)$, $f \in \mathcal{X}$, from noisy observations*

$$\omega = f\big|_{\Gamma_n} + \xi, \xi \sim \mathcal{N}(0; I_n)$$

taken along n-point equidistant grid Γ_n, up to a factor $C(\beta) = [...]$, $\beta := k + \alpha$, is $(Ln^{-\beta})^{1/(2\beta+1)}$, and the upper bound is attained at the affine in ω estimate explicitly given by [...]

Theorem B [Donoho 1994] *Let $\mathcal{X} \subset \mathbb{R}^N$ be a convex compact set, A be an $n \times N$ matrix, and $g(\cdot)$ be a linear form on $\mathcal{X}$. The minimax, over $f \in \mathcal{X}$, risk of recovering $g(f)$ from noisy observations*

$$\omega = Af + \xi, \xi \sim \mathcal{N}(0, I_n),$$

within factor 1.2 is attained at an affine in ω estimate readily given, along with its risk, by the solution to convex optimization problem [...]

♠ **Similarity: A**, **B** are about estimating a linear function of (unknown) "signal" f belonging to a given convex set $\mathcal{X}$ via observation ω of (affine image of) f in white Gaussian noise and claim near minimax optimality of certain efficiently computable affine in ω estimate.

♠ **Difference:**

• **A** is *narrowly focused* (very specific restrictions on $\mathcal{X}$) *descriptive* result – it presents the estimate and its risk in "closed analytic form" ($\Rightarrow$*huge explanation power*). Descriptive results form the bulk of High-Dimensional Statistics and typically are "fragile;" e.g., it is really difficult to extend **A** to the case of *indirect* observations $\omega = Af + \xi$.

• **B** is an *operational* result explaining *how to act* rather than *what to expect*: in **B**, the estimate and its risk are given by *efficient computation* instead of "closed analytic form" expressions ($\Rightarrow$*no explanation power*). **B** is *broadly focused* (all needed is linearity of ω in f and convexity of the set $\mathcal{X}$ of candidate signals) and *guarantees that the computed risk*, whether high or low, *is optimal*, up to 20%, *under the circumstances*.

♣ **Contents of the Talk:** *Near-optimal operational results in hypothesis testing*

♠ **Starting point: Detector-based tests.** Consider the basic problem of deciding on *two composite hypotheses*: *Given two families $\mathcal{P}_1, \mathcal{P}_2$ of probability distributions on a given observation space Ω and an observation $\omega \sim P$ with P known to belong to $\mathcal{P}_1 \cup \mathcal{P}_2$, we want to decide whether $P \in \mathcal{P}_1$ (hypothesis H_1) or $P \in \mathcal{P}_2$ (hypothesis H_2).*

♣ A *detector* is a function $\phi : \Omega \to \mathbb{R}$. *Risks* $\epsilon_{1,2}$, $\epsilon_{2,1}$ of a detector ϕ are defined as

$$\epsilon_{1,2} = \sup_{P \in \mathcal{P}_1} \int_\Omega \mathrm{e}^{-\phi(\omega)} P(d\omega), \;\; \epsilon_{2,1} = \sup_{P \in \mathcal{P}_2} \int_\Omega \mathrm{e}^{\phi(\omega)} P(d\omega)$$

• Given observation $\omega \in \Omega$, *the test $\mathcal{T}_\phi$* associated with detector ϕ accepts H_1 and rejects H_2 when $\phi(\omega) \geq 0$; otherwise the test accepts H_2 and rejects H_1.

♣ **Observation I:** *The probability for $\mathcal{T}_\phi$ to reject the true hypothesis is $\leq \epsilon_{1,2}$ when H_1 is true and is $\leq \epsilon_{2,1}$ when H_2 is true:*

$$\begin{array}{rcl} P \in \mathcal{P}_1 & \Rightarrow & \mathrm{Prob}_{\omega \sim P}\{\omega : \phi(\omega) < 0\} \leq \epsilon_{1,2} \\ P \in \mathcal{P}_2 & \Rightarrow & \mathrm{Prob}_{\omega \sim P}\{\omega : \phi(\omega) \geq 0\} \leq \epsilon_{2,1} \end{array}$$

$$\sup_{P\in\mathcal{P}_1}\int_\Omega e^{-\phi(\omega)}P(d\omega)\leq\epsilon_{1,2},\ \sup_{P\in\mathcal{P}_2}\int_\Omega e^{\phi(\omega)}P(d\omega)\leq\epsilon_{2,1} \qquad (!)$$

♣ **Observation II:** *Detector-based tests admit simple calculus:*

A. *Shift $\phi(\cdot)\mapsto\phi(\cdot)-a$ results in $\epsilon_{1,2}\mapsto\exp\{a\}\epsilon_{1,2}$, $\epsilon_{2,1}\mapsto\exp\{-a\}\epsilon_{2,1}$*
$\Rightarrow$What matters is the product $\epsilon^2:=\epsilon_{1,2}\epsilon_{2,1}$ of the risks: by shift we can redistribute this product between the factors as we wish, e.g., we can make both risks equal to ϵ ("balanced detector")

B. *Detectors are ideally suited to passing from a single observation $\omega\sim P\in\mathcal{P}_1\cup\mathcal{P}_2$ to stationary K -repeated observation – an i.i.d. sample $\omega^K=(\omega_1,...,\omega_K)$ with $\omega_t\sim P$: setting $\phi^{(K)}(\omega^K)=\sum_{t=1}^K\phi(\omega_t)$, the risks of $\phi^{(K)}$ are $\epsilon_{1,2}^{(K)}=\epsilon_{1,2}^K$, $\epsilon_{2,1}^{(K)}=\epsilon_{2,1}^K$.*

C. $(!)$ *is a system of convex constraints on $\phi(\cdot)$, $\epsilon_{1,2}$, $\epsilon_{2,1}$*

D. *P enters $(!)$ linearly $\Rightarrow$risk remains intact when passing from $\mathcal{P}_1$, $\mathcal{P}_2$ to their convex hulls*

E. *Let $\mathcal{T}$ decide on H_1, H_2 with risks $\leq\delta<1/2$. Setting*

$$\phi(\omega)=\tfrac{1}{2}\ln(\delta^{-1}-1)\cdot\left\{\begin{array}{rl}1, & \mathcal{T}\text{ accepts }H_1\\ -1, & \mathcal{T}\text{ accepts }H_2\end{array}\right.,$$

the risks of the resulting detector are $\leq 2\sqrt{\delta(1-\delta)}<1$.

♣ **Conclusion:** *Imagine we can solve the convex optimization problem*

$$\ln(\epsilon_\star)=\frac{1}{2}\min_{\phi(\cdot)}\max_{\substack{P_1\in\mathcal{P}_1\\P_2\in\mathcal{P}_2}}\left[\ln\left(\int_\Omega \mathrm{e}^{\phi(\omega)}P_1(d\omega)\right)+\ln\left(\int_\Omega \mathrm{e}^{-\phi(\omega)}P_2(d\omega)\right)\right]\quad(!)$$

Balanced optimal solution $\phi_\star(\cdot)$ to $(!)$ induces test deciding on H_1, H_2 with risk $\leq\epsilon_\star$ which is near-optimal: whenever H_1, H_2 can be decided upon with risk $\delta<1/2$, it holds

$$\epsilon_\star\leq 2\sqrt{\delta(1-\delta)}.$$

♠ **Difficulty:** *Unless Ω is finite, $(!)$ is an infinite-dimensional problem, and unless $\mathcal{P}_1$, $\mathcal{P}_2$ are finite, $(!)$ is a problem with difficult to compute objective*.

$\Rightarrow$ *In general, $(!)$ is intractable...*

♣ *We are about to consider "good" observation schemes where Difficulty can be circumvented.*

Good Observation Scheme $\mathcal{O} = ((\Omega, P), \{p_\mu : \mu \in \mathcal{M}\}, \mathcal{F})$

♠ (Ω, P): (complete separable metric) *observation space* Ω with (σ-finite σ-additive) *reference measure* P, $\mathrm{supp} P = \Omega$;

♠ $\{p_\mu(\cdot) : \mu \in \mathcal{M}\}$: parametric family of probability densities, taken w.r.t. P, on Ω.

- $\mathcal{M}$ is a relatively open *convex* set in some $\mathbb{R}^n$
- $p_\mu(\omega)$: *positive* and continuous in $\mu \in \mathcal{M}, \omega \in \Omega$

♠ $\mathcal{F}$: *finite-dimensional* space of continuous functions on Ω containing constants and such that

$$\ln(p_\mu(\cdot)/p_\nu(\cdot)) \in \mathcal{F} \ \forall \mu, \nu \in \mathcal{M}$$

♠ *For $\phi \in \mathcal{F}$, the function* $\mu \mapsto \ln\left(\int_\Omega e^{\phi(\omega)} p_\mu(\omega) P(d\omega)\right)$ *is finite and concave in $\mu \in \mathcal{M}$.*

♣ **Example I: Gaussian o.s.**

$$\left((\Omega = \mathbb{R}^d, d\omega), \{p_\mu(\cdot) = \mathcal{N}(\mu, I_d) : \mu \in \mathbb{R}^d\}, \mathcal{F} = \{a^T\omega + b\}_{\substack{a\in\mathbb{R}^d\\ b\in\mathbb{R}}}\right)$$

$\Rightarrow \ln\left(\int e^{a^T\omega+b} p_\mu(\omega)d\omega\right) = \frac{a^Ta}{2} - a^T\mu + b$ indeed is concave in μ

♣ **Example II: Poisson o.s.** Here (Ω, P) is $\mathbb{Z}^d_+$ with counting measure,

$$\mathcal{M} = \mathbb{R}^d_{++},\ p_\mu(\omega) = \prod_{i=1}^d \left[\frac{\mu_i^{\omega_i}}{\omega_i!} e^{-\mu_i}\right]$$

and $\mathcal{F}$ is the family of affine functions on Ω

$\Rightarrow \ln\left(\int e^{a^T\omega+b} p_\mu(\omega)d\omega\right) = b + \sum_{i=1}^d \mu_i[e_i^a - 1]$ indeed is concave in μ.

♣ **Example III: Discrete o.s.** Here (Ω, P) is a finite set $\{1, ..., d\}$ with counting measure,

$$\mathcal{M} = \{\mu \in \mathbb{R}^d_{++} : \textstyle\sum_{\omega=1}^d \mu_\omega = 1\},\ p_\mu(\omega) = \mu_\omega,\ \omega \in \Omega$$

and $\mathcal{F}$ is the family of all functions on Ω

$\Rightarrow \ln\left(\int e^{\phi(\omega)} p_\mu(\omega)d\omega\right) = \ln\left(\sum_{\omega=1}^d e^{\phi(\omega)}\mu_\omega\right)$ indeed is concave in μ.

♣ **Example IV: Direct Product of good o.s.'s:** Given K good o.s.'s

$$\mathcal{O}_t = ((\Omega_t, P_t), \{p_{\mu_t,t}(\cdot) : \mu_t \in \mathcal{M}_t\}, \mathcal{F}_t), 1 \leq t \leq K$$

their *direct product* is defined as

$$\begin{array}{rcl}\mathcal{O}^{(K)} & = & \Big((\Omega^{(K)} = \bigotimes_{t=1}^{K} \Omega_t, P^{(K)} = \bigotimes_{t=1}^{K} P_t), \\ && \{p_\mu(\omega^K) = \prod_{t=1}^{K} p_{\mu_t}(\omega_t) : \mu = [\mu_1; ...; \mu_K] \in \mathcal{M}^{(K)} = \bigotimes_{t=1}^{K} \mathcal{M}_t\}, \\ && \mathcal{F}^{(K)} = \{f(\omega^K) = \sum_{t=1}^{K} f_t(\omega_t) : f_t \in \mathcal{F}_t\}\Big)\end{array}$$

and describes a sample of *independent* observations $\omega^K = (\omega_1, ..., \omega_K)$ with ω_t drawn from $\mathcal{O}_t$. *Direct product of good o.s.'s is good.*

♣ When all factors $\mathcal{O}_t = \mathcal{O} := ((\Omega, P), \{p_\mu : \mu \in \mathcal{M}\}, \mathcal{F})$ are identical, we can "restrict direct product on diagonal", arriving at a good o.s.

$$\begin{array}{rcl}\mathcal{O}^K & = & \Big((\Omega^K = \bigotimes_{t=1}^{K} \Omega, P^K = \bigotimes_{t=1}^{K} P), \\ && \{p_{\mu,K}(\omega^K) = \prod_{t=1}^{K} p_\mu(\omega_t) : \mu \in \mathcal{M}\}, \mathcal{F}^K = \{f(\omega_t) = \sum_{t=1}^{K} f(\omega_t) : f \in \mathcal{F}\}\Big)\end{array}$$

representing *stationary K-repeated observations* $\omega^K = (\omega_1, ..., \omega_K)$ with i.i.d. ω_t drawn from $\mathcal{O}$.

$$\ln(\epsilon_\star) = \frac{1}{2} \min_{\phi(\cdot)} \max_{\substack{P_1 \in \mathcal{P}_1 \\ P_2 \in \mathcal{P}_2}} \left[\ln\left(\int_\Omega e^{\phi(\omega)} P_1(d\omega) \right) + \ln\left(\int_\Omega e^{-\phi(\omega)} P_2(d\omega) \right) \right] \quad (!)$$

♠ **Main Theorem:** *Let* $\mathcal{O} := ((\Omega, P), \{p_\mu : \mu \in \mathcal{M}\}, \mathcal{F})$ *be a good o.s., and let*

$$\mathcal{P}_1 = \{p_\mu(\omega)P(d\omega) : \mu \in X_1\},\ \mathcal{P}_2 = \{p_\mu(\omega)P(d\omega) : \mu \in X_2\}$$

where X_1, X_2 *are nonempty convex compact subsets of* $\mathcal{M}$. *The optimization problem*

$$\ln(\epsilon_\star) = \max_{\substack{\mu \in \mathcal{X}_1 \\ \nu \in \mathcal{X}_2}} \ln\left(\int_\Omega \sqrt{p_\mu(\omega) p_\nu(\omega)} P(d\omega) \right)$$

is convex and solvable. Given optimal solution (μ_*, ν_*) *to the problem, the detector*

$$\phi_*(\omega) = \frac{1}{2} \ln(p_{\mu_*}(\omega)/p_{\nu_*}(\omega))$$

is a balanced optimal solution to problem $(!)$. *Consequently, the test given by the detector* ϕ_* *decides near optimally upon the hypotheses*

$$H_1 : \omega \sim p_\mu(\cdot) \text{ with } \mu \in X_1 \text{ and } H_2 : \omega \sim p_\mu(\cdot) \text{ with } \mu \in X_2$$

♣ Let us apply Main Theorem to the stationary K-repeated version $\mathcal{O}^K$ of a good o.s. $\mathcal{O} := ((\Omega, P), \{p_\mu : \mu \in \mathcal{M}\}, \mathcal{F})$. The associated optimization problem is

$$\begin{array}{rcl}\ln(\epsilon_\star^{(K)}) & = & \max\limits_{\substack{\mu\in\mathcal{X}_1\\ \nu\in\mathcal{X}_2}} \ln\left(\int\limits_{\Omega^K} \sqrt{\prod\limits_{t=1}^K (p_\mu(\omega_t)p_\nu(\omega_t))} P(d\omega_1)...P(d\omega_K)\right) \\ & = & \max\limits_{\substack{\mu\in\mathcal{X}_1\\ \nu\in\mathcal{X}_2}} K \ln\left(\int\limits_{\Omega} \sqrt{p_\mu(\omega)p_\nu(\omega)} P(d\omega)\right)\end{array}$$

$\Rightarrow$ *The optimal solution (μ_*, ν_*) to the problem is the same as the optimal solution to the single-observation problem*

$$\ln(\epsilon_\star) = \max_{\substack{\mu\in\mathcal{X}_1\\ \nu\in\mathcal{X}_2}} \ln\left(\int\limits_{\Omega} \sqrt{p_\mu(\omega)p_\nu(\omega)} P(d\omega)\right),$$

one has

$$\epsilon_\star^{(K)} = \epsilon_\star^K$$

and the detector based on K-repeated observations is

$$\phi^K(\omega^K) = \sum_{t=1}^K \phi(\omega_t), \ \phi(\cdot) = \frac{1}{2}\ln(p_{\mu_*}(\cdot)/p_{\nu_*}(\cdot)).$$

$\Rightarrow$The near-optimality claim of Main Theorem can be reformulated as follows:

♠ *Let* $\mathcal{O} = ((\Omega, P), \{p_\mu : \mu \in \mathcal{M}\}, \mathcal{F})$ *be a good o.s.,* X_1, X_2 *be convex compact subsets of* $\mathcal{M}$*, and* $\bar{K}$ *be a positive integer. Assume that in the nature there exists a test based on* $\bar{K}$*-repeated observations deciding upon the hypotheses*

$$H_1 : \omega \sim p_\mu(\cdot) \text{ with } \mu \in X_1 \text{ and } H_2 : \omega \sim p_\mu(\cdot) \text{ with } \mu \in X_2$$

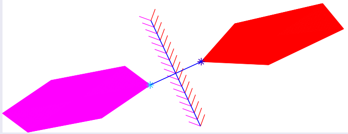
with risk $\leq \epsilon < 1/4$*. Then the test yielded by the Main Theorem as applied to* X_1, X_2 *and* K*-repeated version of* $\mathcal{O}$ *decides on* H_1*,* H_2 *with risk* $\leq (2\sqrt{\epsilon})^{K/\bar{K}}$*. The latter risk is* $\leq \epsilon$ *provided that*

$$K \geq \left[\frac{2\ln(1/\epsilon)}{\ln(1/\epsilon) - 2\ln(2)}\right] \bar{K}.$$

Generic Application, I. Gaussian Signal Processing:

$$\omega = \mathcal{A}(u) + \xi,\ \xi \sim \mathcal{N}(0, I_d) \text{ with affine } \mathcal{A}(\cdot) : \mathbb{R}^n \to \mathbb{R}^d$$

- We can decide upon hypotheses H_1 and H_2 stating, respectively, that $u \in U$ and $u \in V$ with convex compact sets U, V. The test deciding on H_1 vs. H_2 is as follows:
    - We solve the convex program $\mathrm{Opt} = \min\limits_{u \in U, v \in V} \frac{1}{8}\|\mathcal{A}(u) - \mathcal{A}(v)\|_2^2$. An optimal solution u_*, v_* defines detector $\phi_*(\omega) = \left[\frac{\mathcal{A}(u_*) - \mathcal{A}(v_*)}{2}\right]^T \left[\omega - \frac{\mathcal{A}(u_*) + \mathcal{A}(v_*)}{2}\right]$



Magenta polygon: $\mathcal{A}(U)$ Red polygon: $\mathcal{A}(V)$

- An *upper risk bound*, as given by our theory, is $\epsilon_\star = \mathrm{e}^{-\mathrm{Opt}}$. The *actual risk* of the test is $\mathrm{Erf}(r) := \frac{1}{\sqrt{2\pi}} \int\limits_r^\infty \mathrm{e}^{-s^2/2} ds$, $r = \frac{1}{2}\|\mathcal{A}(u_*) - \mathcal{A}(v_*)\|_2$, and the test is *optimal* in risk.

Note: In the Gaussian case, the optimal test is self-evident and can be built without any science.

Generic Application, II. Poisson Imaging:

$$\omega = [\omega_1; ...; \omega_d],\ \omega_s = \mathrm{Poisson}([Au]_s) \text{ independent across } s = 1, ..., d$$
$$\left[A \in \mathbb{R}_+^{d\times n}, u \in \mathbb{R}_+^n\right]$$

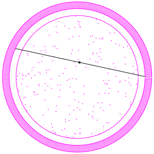
• We can decide upon hypotheses H_1 and H_2 stating, respectively, that $u \in U$ and $u \in V$, with convex compact sets $U \subset \mathbb{R}_+^n$, $V \subset \mathbb{R}_+^n$.

Note: The Poisson model is responsible for

- Positron Emission Tomography
- Large Binocular Telescope – cutting edge astronomical imaging instrument under development by an international consortium
- Nanoscale Fluorescent Microscopy (Poisson Biophotonics) – a revolutionary technology allowing to break the diffraction barrier and to view biological molecules "at work" at a resolution 10-20 nm, yielding entirely new insights into the signalling and transport processes within cells.

♣ **Example: Positron Emission Tomography.** In PET, a patient is injected radioactive tracer and placed inside a cylinder with the surface split into *detector cells*.

• Every act of tracer's disintegration yields two γ-quants flying in opposite directions along a randomly oriented *line of response* (l.o.r.). Eventually the quants "simultaneously" (within time window like 10^{-8} sec) hit two detector cells ("coincidence" which is registered).



Ring of detector cells and line of response

• Observation is the list of total numbers of coincidences registered in every *bin* (pair of detector cells) over a given time T, and the goal is to make inferences about density u of the tracer. After discretization, we arrive at Poisson o.s.

$$\omega = \{\omega_s \sim \mathrm{Poisson}(\textstyle\sum_{j=1}^n A_{sj}u_j)\}_{s=1}^d$$

• d: # of bins • n: # of *voxels* (small 3D cubes in which the field of view is split)
• u_j: average tracer's density in voxel j • A_{sj}/T: probability for l.o.r. originating in voxel j to be registered in bin s

♠ In Poisson case $\omega=\{\omega_s\sim \mathrm{Poisson}([\mathcal{A}(u)]_s)\}_{s=1}^d$, the test deciding on $u\in U$ vs. $u\in V$ is as follows:

- We solve the convex program

$$\mathrm{Opt}=\min_{u\in U,v\in V}\frac{1}{2}\sum_{s=1}^d\left[\sqrt{[\mathcal{A}(u)]_s}-\sqrt{[\mathcal{A}(v)]_s}\right]^2.$$

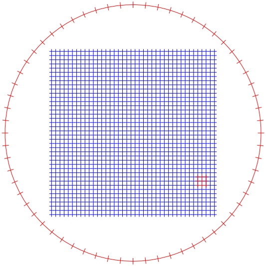
- An optimal solution u_*,v_* defines the detector

$$\phi_*(\omega) \quad = \quad \frac{1}{2}\sum_{s=1}^d \ln\left(\sqrt{[\mathcal{A}(u_*)]_s/[\mathcal{A}(v_*)]_s}\right)\omega_s-\frac{1}{2}\sum_{s=1}^d[\mathcal{A}(u_*)-\mathcal{A}(v_*)]_s.$$

- The *upper risk bound* is $\epsilon_\star=\mathrm{e}^{-\mathrm{Opt}}$.

How It Works: PET

♣ **Illustration:** We consider 2D PET with $m = 64$ detector cells and 40×40 field of view:

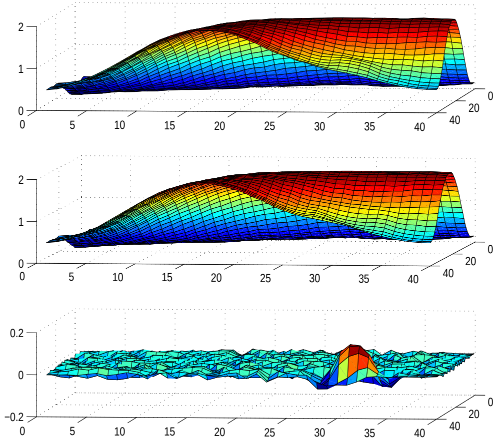


Detector cells and field of view. 1296 bins, 1600 pixels

- $\mathcal{W}$: the set of tracer's densities $w \in \mathbb{R}^{40\times 40}$ satisfying some regularity assumptions and at average not exceeding 1
- U: densities from $\mathcal{W}$ with the average over the 3×3 red spot at least 1.1
- V: densities from $\mathcal{W}$ with average over the red spot at most 1.
- The observation time is chosen to allow to decide on H_1 vs. H_2 with risk 0.01.

♠ Simulation results:

- *In 1024 simulations where H_1 was true, the hypothesis was rejected in 0% of cases*
- *In 1024 simulations where H_2 was true, the hypothesis was rejected in 0.1% of cases*



- Top: u_* • Middle: v_* • Bottom: $u_* - v_*$

Testing Multiple Hypotheses

♣ **Situation:** We are given good o.s. $\mathcal{O} = ((\Omega, P), \{p_\mu : \mu \in \mathcal{M}\}, \mathcal{F})$ and M hypotheses on the distribution of observation ω:

$$H_i : \omega \sim p_\mu(\omega) \text{ with } \mu \in X_i,\ 1 \leq i \leq M$$

where X_i are *convex compact* subsets of $\mathcal{M}$ ("convex hypotheses"). Given stationary K-repeated observation ω^K, we want to decide on the hypotheses H_1,...,H_M.

♠ **Note:** We do *not* insist on "full separation" of the hypotheses. Instead, we are given a *symmetric 0/1 "closeness matrix"* $\mathcal{C} = [\mathcal{C}_{ij}]_{\substack{1\leq i\leq M\\ 1\leq j\leq M}}$ *with zero diagonal*, where

- $\mathcal{C}_{ij} = 0$ means *"H_i and H_j are close to each other"*
- $\mathcal{C}_{ij} = 1$ means *"H_i and H_j are far from each other"*

and we do *not* insist on distinguishing close to each other hypotheses.

♣ **Typical application: Inferring Color.** Assume that every X_i is assigned *color* $c_i \in \{1, ..., m\}$, and let the sets of different colors do not intersect. Our goal is to infer from $\omega^K \sim p_{\mu,K}(\cdot)$ with unknown $\mu \in \bigcup\limits_{i=1}^{M} X_i$ the color of μ (i.e., the common color of all X_i containing μ). Setting $\mathcal{C}_{ij} = 0$ iff $c_i = c_j$, *Inferring Color* reduces to deciding upon H_1,..., H_M "up to closeness."

♠ **Strategy:**

• [**Building basic detectors**] Using Main Theorem, we build balanced pairwise detectors $\phi_{ij}(\cdot)$ with risks ϵ_{ij}, $1 \leq i < j \leq M$ underlying near-optimal tests deciding on H_i vs. H_j, and let us set

$$\phi_{ji}(\omega) \equiv -\phi_{ij}(\omega),\ \epsilon_{ji} = \epsilon_{ij},\ 1 \leq i < j \leq M,$$
$$\phi_{ii}(\omega) \equiv 0,\ \epsilon_{ii} = 1,\ 1 \leq i \leq M$$
$$\phi_{ij}^K(\omega^K) = \sum_{t=1}^K \phi_{ij}(\omega_t)$$

thus arriving at

$$\phi_{ij}^K \equiv -\phi_{ji}^K,\ \epsilon_{ij} = \epsilon_{ji},\ 1 \leq i, j \leq M$$
$$\int_\Omega \mathrm{e}^{-\phi_{ij}^K(\omega^K)} p_{\mu,K}(\omega^K) P^K(d\omega^K) \leq \epsilon_{ij}^K\ \forall \mu \in X_i\, \forall i, j$$

• [**Test**] Let us select a skew-symmetric scaling matrix $\alpha = [\alpha_{ij}]_{\substack{1\leq i\leq M\\ 1\leq j\leq M}}$ and consider the test $\mathcal{T}^K$ which, given ω^K, accepts all hypotheses H_i such that

$$\phi_{ij}^K(\omega^K) > \alpha_{ij}\ \forall (j:\ H_j \text{ is far from } H_i)$$

Note: $\mathcal{T}^K$ can accept several (e.g., none of) hypotheses.

$$\phi_{ij}^K \equiv -\phi_{ji}^K,\ \epsilon_{ij} = \epsilon_{ji},\ 1 \le i,j \le M$$
$$\int_\Omega \mathrm{e}^{-\phi_{ij}^K(\omega^K)} p_{\mu,K}(\omega^K) P^K(d\omega^K) \le \epsilon_{ij}^K\ \forall \mu \in X_i\, \forall i,j$$

- Given ω^K, $\mathcal{T}^K$ accepts H_i iff $\phi_{ij}^K(\omega^K) > \alpha_{ij}\ \forall (j:\ H_j$ is far from $H_i)$

♣ **Theorem:** *Let* $\epsilon_\star^{(K)} = \max\limits_{1\le i\le M} \sum\limits_{j:\mathcal{C}_{ij}=1} \epsilon_{ij}^K \mathrm{e}^{\alpha_{ij}}$. *The risk of* $\mathcal{T}^K$ *is at most* $\epsilon_\star^{(K)}$*, meaning that whenever* ω^K *is drawn from* $p_{\mu,K}(\cdot)$ *with* $\mu \in X_{i_*}$ *and* $\mathcal{T}^K$ *is applied to* ω^K*, the* $p_{\mu,K}$*-probability of the event*

> *the true hypothesis* H_{i_*} *is accepted and all other accepted hypotheses are close to* H_{i_*}

is at least $1 - \epsilon_\star^{(K)}$.

Note: The smallest, over skew-symmetric matrices α, value of the risk $\epsilon_\star^{(K)}$ is the spectral norm (or, which is the same, the Perron-Frobenius eigenvalue) of the symmetric nonnegative matrix

$$E = \left[\epsilon_{ij}^K \mathcal{C}_{ij}\right]_{\substack{1\le i\le M\\ 1\le j\le M}}$$

♣ **Theorem** [near-optimality] *Assume that for some $\bar{K}$ "in the nature" there exists a test, based on stationary $\bar{K}$-repeated observations, which decides upon the hypotheses $H_1, ..., H_M$ with some risk $\epsilon < 1/4$. Whenever*

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

the risk of $\mathcal{T}^K$ is $\leq \epsilon$.

Application: Estimating L-convex functional on a union of convex sets

♣ **Situation:** Given are:

- a good o.s. $\mathcal{O} = ((\Omega, P), \{p_\mu : \mu \in \mathcal{M}\}, \mathcal{F})$
- a convex compact set $X \subset \mathcal{M}$ and M closed convex sets $X_i \subset X$
- a continuous function $f(\cdot) : X \to \mathbb{R}$ which is *L-convex*, meaning that *the sets $\{x \in X : f(x) \le a\}$ and $\{x \in X : f(x) \ge a\}$ are unions of at most L convex sets.*

Example: affine-fractional function with no singularities on X is 1-convex.

• **Goal:** To estimate $f(\mu_*)$ with *unknown* μ_* known to belong to $\mathcal{X} := \bigcup_{i=1}^{M} X_i$ via stationary K-repeated observation $\omega^K \sim p_{\mu_*,K}(\cdot)$.

♣ **Observation:** Given $\ell < u$, the sets $\{\mu \in \mathcal{X} : f(\mu) \le \ell\}$, $\{\nu \in \mathcal{X} : f(\mu) \ge u\}$ are unions of at most LM convex compact sets. Coloring the resulting $2LM$ sets in magenta and red according to their origin, the Inferring Color procedure

- recognizes correctly, up to the risk of the procedure, that $f(\mu_*) \le \ell$ or $f(\mu_*) \ge u$, if these are the cases
- is unpredictable when $\ell < f(\mu_*) < u$.

♠ **Bisection:** Given a *localizer* $\Delta = [a, b]$ presumably containing $f(\mu_*)$,
• set $c = \frac{a+b}{2}$ and find $r < c$, $s > c$ as close to c as possible under the restriction that *the risk of wrong color inference in every one of the Color Inferring problems*

- $\{\mu \in \mathcal{X} : f(\mu) \le r\}$ *vs.* $\{\mu \in \mathcal{X} : f(\mu) \ge c\}$
- $\{\mu \in \mathcal{X} : f(\mu) \le c\}$ *vs.* $\{\mu \in \mathcal{X} : f(\mu) \ge s\}$

is at most a given ϵ.
• terminate if either $c - r > \frac{1}{4}[b - a]$, or $s - c > \frac{1}{4}[b - a]$, or both, otherwise
• use ω^K in Inferring Color to decide on the color of μ_* in the first and in the second problem. Take as a new localizer the segment

$[a, s]$, if both times the inferred color of μ_* is magenta
$[r, b]$, if both times the inferred color of μ_* is red
$[r, s]$, if the two inferred colors of μ_* differ from each other

In the first two cases, pass to the next Bisection step, in the third terminate.

• **Note:**

- *Every Bisection step reduces the size of localizer by at least 1/4*
- *With initial localizer covering* $f(\mu_*)$*, the probability for the first* $N = 1, 2, ...$ *localizers to cover* $f(\mu_*)$ *is at least* $1 - 2N\epsilon$
- *The procedure is minimax optimal within logarithmic in M, L factors.*

Application: Sequential Hypothesis Testing

♣ **Situation:** Given are

- a good o.s. $\mathcal{O} = ((\Omega, P), \{p_\mu : \mu \in \mathcal{M}\}, \mathcal{F})$
- convex compact sets $X_i \subset \mathcal{M}$, $1 \le i \le M$, colored in m colors.

♣ **Assumption:** Sets of different colors are at positive Hellinger distance from each other.

♣ **Goal:** To infer, with a given risk ϵ, the color of p_μ, $\mu \in \mathcal{X} := \bigcup\limits_{i=1}^{M} X_i$, from stationary K-repeated observation $\omega^K = (\omega_1, ..., \omega_K) \sim p_{\mu,K}(\cdot)$.

♠ By our theory, the risk of Inferring Color via ω^K is the spectral norm of the matrix $E^{(K)} = [\epsilon^K_{X_i:X_j} \mathcal{C}_{ij}]_{\substack{1 \le i \le M \\ 1 \le j \le M}}$ where $\mathcal{C}_{ij}$ is 0 or 1 depending on whether the colors of i, j are the same or different, and

$$\epsilon_{X_i:X_j} = \max_{\mu \in X_i, \nu \in X_j} \int\limits_\Omega \sqrt{p_\mu(\omega) p_\nu(\omega)} P(d\omega).$$

- The sets X_i, X_j of different colors are at positive Hellinger distance

$\Rightarrow \epsilon_{X_i:X_j} \mathcal{C}_{ij} < 1$ for all i, j

$\Rightarrow$ *For large K, the risk of our near-optimal Inferring Color procedure is* $\le \epsilon$.

♠ **However:** The required value of K is governed by the *closest to 1* of the quantities $\epsilon_{ij}\mathcal{C}_{ij}$ and can be large if the Hellinger distance between some pair of sets X_i, X_j of different colors is small.

♠ **Question:** *Can we process observations $\omega_t \sim p_\mu(\cdot)$, $t = 1, 2, ..., K$, one by one and to decide on the color of μ on-line, in order to make rapid decisions when μ is "deeply inside" one of the sets X_i ?*

♣ **Sketch of Strategy:**

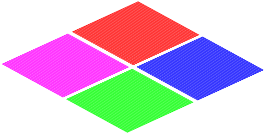
• Let us represent, dynamically in time t, every X_i as the union of a number of convex compact sets X^t_{ij}, with one of the sets, X^t_{i0}, covering the *growing with time* "inner part" of X_i, and remaining sets covering a *shrinking with time* stripe $X_i \backslash X^t_{i0}$. As a result, we get convex compact sets X^t_i, $1 \leq i \leq M_t$, colored in m colors.

• At every time t, we build the Inferring Color procedure for the hypotheses $H^t_i : \mu \in X^t_i$, $1 \leq i \leq M_t$, defining the closeness $\mathcal{C}^{(t)}$ by $\mathcal{C}^{(t)}_{ij} = 0$ *iff* $\epsilon_{X^t_i : X^t_j} > 1 - \delta_t$, with δ_t as small as possible under the restriction

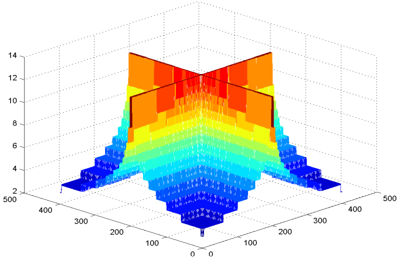
$$\|[\epsilon^t_{X^t_i : X^t_j} \mathcal{C}^{(t)}_{ij}]_{i,j}\|_{2,2} \leq \epsilon_t \qquad [\epsilon_t > 0 : \textstyle\sum_t \epsilon_t = \epsilon]$$

• We apply the Inferring Color procedure to ω^t. If some of the hypotheses H^t_i are accepted and *all accepted hypotheses are of the same color*, we claim that μ is of this color and terminate, otherwise proceed to the next observation ω_{t+1}.

♠ With proper implementation, the resulting Sequential Hypotheses Testing is near-optimal in the minimax sense *and indeed results in significant savings in observation time when the distribution underlying the observations is "deeply inside" the set of distributions of the same color:*

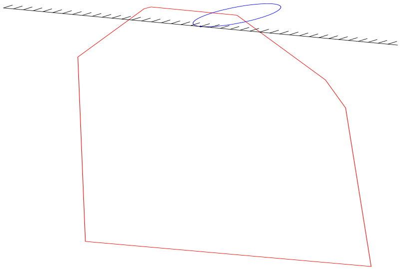


Four squares X_1, X_2, X_3, X_4
Gaussian o.s.



Logarithm of observation time
times: max$= 1.6 \cdot 10^5$, median$= 154$

♣ **Induced problem:** *Given two compact convex sets X, Y, how to build a convex set Z such that $Z \cap Y = \emptyset$ and $X \backslash Z$ is "as small as possible" ?*



Red set: X; Blue set: Y

Goal: "to minimize" the intersection of the

black half-space and the red set

Dynamical Hypothesis Testing

♣ In Sequential Hypothesis Testing, the hypotheses are stationary, and the observations are i.i.d.
In *Dynamical Hypothesis Testing*, the hypotheses "evolve in time."

♠ **Model:** We observe on time horizon $1, ..., T$ noisy output of a known discrete time linear system with "memory depth" W:

$$\omega_t = \sum_{\tau=0}^{W-1} a_\tau u_{t-\tau} + \xi_t, \quad 1 \le t \le T$$

- $u = \{u_t\}_{t=-W+2}^{T} \in \mathbb{R}^{W+T-1}$: input
- $x[u] = \{x_t[u] := \sum_{\tau=0}^{W-1} a_\tau u_{t-\tau}\}_{t=1}^{T}$: output
- $\xi = [\xi_1; ...; \xi_T] \sim \mathcal{N}(0, I_T)$: observation noise

♣ **Given are:** tolerance $\epsilon \in (0,1)$ and M hypotheses

- H_1: $u \in U_1$ [nuisance input]
- H_i: $u \in U_i$, $2 \le i \le M$ [signal input]

where U_i are closed convex sets in $\mathbb{R}^{W+T-1}$.

♣ **Goal:** under the restriction that *the probability of false alarm* – qualifying a nuisance input as a non-nuisance – *does not exceed* ϵ, *to recognize as early as possible that the actual input is a non-nuisance.*

Example: Nuisance hypothesis H_1 states that the input is zero ($U_1 = \{0\}$), while signal hypotheses H_j, $j \geq 2$, state that the input is a step starting at specific time instant with the value of (at least) a given magnitude, so that U_i, $i \leq 2$, are obtained by arranging into a sequence the sets

$$U_{t,\chi,\rho} = \left\{ u : u_\tau = \left\{ \begin{array}{ll} 0, & \tau < t \\ \chi\rho, & \tau \geq t \end{array} \right. \right\},$$

where $-W + 2 \leq t \leq T$, $\chi = \pm 1$ and ρ runs through some finite grid on the positive ray.

♣ **Strategy:**

- For $t \in \{1, ..., T\}$, let $X_i^t = \{\{x_\tau[u]\}_{\tau=1}^t : u \in U_i\}$, and let H_i^t, $1 \le i \le M$, state that the observation $\omega^t = (\omega_1, ..., \omega_t)$ is a point from X_i^t corrupted by white Gaussian noise $\mathcal{N}(0, I_t)$.
- The sets X_i^t are convex $\Rightarrow$we can apply Main Theorem to build pairwise detectors $\phi_{ijt}(\cdot)$ and risks ϵ_{ijt} associated with the sets X_i^t:

$$\phi_{ijt} \equiv -\phi_{jit},\ \epsilon_{ijt} = \epsilon_{jit},\ \frac{1}{(2\pi)^{t/2}} \int \mathrm{e}^{-\phi_{ijt}(\omega^t)} \exp\{-\|\omega^t\|_2^2/2\} d\omega^t \le \epsilon_{ijt}.$$

- Define closeness $\mathcal{C}^t$ as follows:
  - *all signal hypotheses H_i^t, $i \ge 2$, are close to each other*
  - *nuisance hypothesis H_1^t is close to a signal hypothesis H_j^t iff* $\epsilon_{1jt} > \delta_t$
    - δ_t: as large as possible under the restriction $\|[\epsilon_{ijt}\mathcal{C}_{ij}^t]_{i,j}\|_{2,2} \le \epsilon/T$.
- Apply to ω^t the multiple hypotheses test corresponding to the resulting closeness and ϵ_{ijt}. *If the test rejects H_1^t, claim that the input is non-nuisance and terminate, otherwise claim that so far the nuisance hypothesis holds true and pass to the next observation, if any.*

♣ On a close inspection, the resulting procedure is as follows:

1. For $j = 2, 3, ..., M$, find $(x_*, y_*) \in \mathrm{Argmin}_{x \in X_1^t, y \in X_j^t} \|x - y\|_2^2$ and set

$$\epsilon_{1jt} = \mathrm{e}^{-\|x_* - y_*\|_2^2/8}, \ \phi_{1jt}(\omega^t) = \frac{1}{2}[x_* - y_*]^T \omega^t + \frac{1}{4}[\|y_*\|_2^2 - \|x_*\|_2^2].$$

Assume w.l.o.g. that $\epsilon_{12t} \geq \epsilon_{13t} \geq ... \geq \epsilon_{1Mt}$.

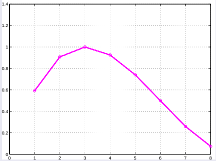
2. Find the smallest $k = k_t$ such that

$$\delta_k := \sqrt{\sum_{j>k} \epsilon_{1jt}^2} \leq \epsilon / T$$

3. *If* $\phi_{1jt}(\omega^t) < \ln(\epsilon_{1jt}/\delta_{k_t})$ *for some* $j > k_t$*, claim that the input is non-nuisance and terminate, otherwise claim that so far the nuisance hypothesis holds true and pass to the next observation, if any.*

How It Works

- Observation horizon $1, 2, ..., 16$, memory depth 8



Impulse Response

- Nuisance hypothesis H_1: input identically zero.
- Signal hypotheses H_i, $i \geq 2$: 1748 hypotheses of the form

input x satisfies $x_\tau = 0$, $\tau < t$, and $\chi x_\tau \geq r$, $\tau \geq t$

with $-6 \leq t \leq 16$, $\chi = \pm 1$ and $r \in \{1.1^s : -15 \leq s \leq 22\}$

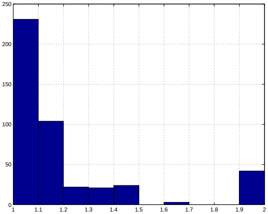
- Probability of false alarm $\epsilon = 0.01$

♠ **Performance metrics:**

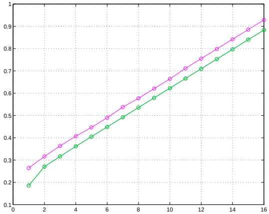
• Signal hypothesis is 0.99-visible at time t, if with the minimal, over inputs obeying the hypothesis, $\|\cdot\|_2$-norm of the output restricted onto $1, ..., t$ is at least $2\mathrm{ErfInv}(0.99)$
Among 506 signal hypotheses, there are 470 visible at time $T = 16$ or earlier.

• Signal hypothesis is 0.99-recognizable at time t, if when the hypothesis is true, the test with probability ≥ 0.99 terminates at time $\leq t$ with "no-nuisance" conclusion.
Among 470 visible hypotheses there are 447 recognizable at time $T = 16$ or earlier (95%).

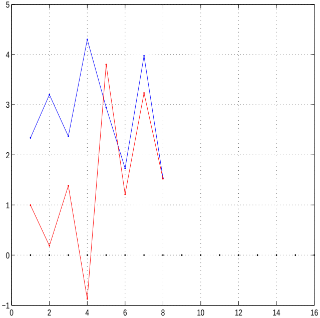
• *Delay* of a 0.99-recognizable somewhere on the entire time horizon signal hypothesis is the ratio of the first instant when the hypothesis becomes 0.99-recognizable to the first instant when it becomes 0.99-visible.



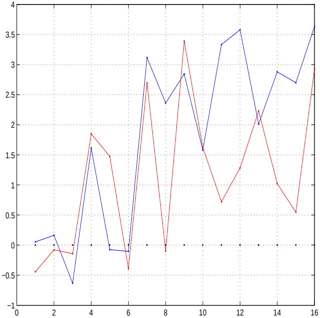
Histogram of delays
min: 1.00 max: 2.00
median: 1.10 mean: 1.20



Percentage of
0.99-visible/0.99-recognizable
hypotheses at time t



Who is nuisance?



Who is nuisance?