Hypothesis Testing via Convex Optimization

Arkadi Nemirovski

Joint research with

Alexander Goldenshluger
Haifa University

Anatoli Iouditski
Grenoble University

Information Theory, Learning and Big Data Workshop
Simons Institute for the Theory of Computing
Berkeley, March 2015

♠ Compare 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 $x(0)$, $x \in \mathcal{X}$, 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 $x \in \mathcal{X}$, risk of recovering $g(x)$ from noisy observations*

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

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" x from a given convex set $\mathcal{X}$ via observation ω of (affine image of) x in white Gaussian noise. Both **A**, **B** claim near minimax optimality of certain efficiently computable affine in ω estimate.

♠ **Difference:**

• **A** is *narrowly focused* (very special $\mathcal{X}$) *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 $\omega = Ax + \xi$.

• **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* instebad of "closed analytic form" expressions ($\Rightarrow$*no explanation power*). **B** is *broadly focused* (all needed is linearity of ω in x 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 $\mathbf{H}_1$) or $P \in \mathcal{P}_2$ (hypothesis $\mathbf{H}_2$).*

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

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

• Given observation $\omega \in \Omega$, *the test $\mathcal{T}_\phi$* associated with detector ϕ accepts $\mathbf{H}_1$ and rejects $\mathbf{H}_2$ when $\phi(\omega) \geq 0$; otherwise the test accepts $\mathbf{H}_2$ and rejects $\mathbf{H}_1$.

♣ **Observation I:** *The probability for $\mathcal{T}_\phi$ to reject the true hypothesis is $\leq \epsilon_{1,2}$ when $\mathbf{H}_1$ is true and is $\leq \epsilon_{2,1}$ when $\mathbf{H}_2$ 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:*

♠ *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* ϵ^2 *between the factors as we wish, e.g., we can make both risks equal to* ϵ *("balanced detector")*

♠ *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* $\epsilon_{1,2}^{(K)}=\epsilon_{1,2}^K$, $\epsilon_{2,1}^{(K)}=\epsilon_{2,1}^K$.

♠ $(!)$ *is a system of convex constraints on* $\phi(\cdot)$, $\epsilon_{1,2}$, $\epsilon_{2,1}$
♠ P *enters* $(!)$ *linearly* $\Rightarrow$ *risk remains intact when passing from* $\mathcal{P}_1$, $\mathcal{P}_2$ *to their convex hulls*

♠ *Let* $\mathcal{T}$ *decide on* $\mathbf{H}_1$, $\mathbf{H}_2$ *with risks* $\leq\delta<1/2$. *Setting*

$$\phi(\omega)=\frac{1}{2}\ln(\delta^{-1}-1)\cdot\begin{cases}1, & \mathcal{T}\text{ accepts }\mathbf{H}_1\\ -1, & \mathcal{T}\text{ accepts }\mathbf{H}_2\end{cases},$$

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

♣ Conclusion:

Imagine we can solve the convex optimization problem

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

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...

♣ This talk:

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 $\mu \in \mathcal{M}$.*

Gaussian o.s.

$(\Omega = \mathbb{R}^d, d\omega)$, $\{p_\mu(\cdot) = \mathcal{N}(\mu, I_d) : \mu \in \mathbb{R}^d\}$, $\mathcal{F} = \{\text{affine functions on } \Omega\}$

Poisson o.s.

$(\Omega, P) = (\mathbb{Z}_+^d, \text{counting measure})$, $\{p_\mu(\omega) = \prod_{i=1}^{d} \frac{\mu_i^{\omega_i} e^{-\mu_i}}{\omega_i!} : \mu \in \mathcal{M} := \mathbb{R}_{++}^d\}$,
$\mathcal{F} = \{\text{affine functions on } \Omega\}$

Discrete o.s.

$(\Omega, P) = (\{1, ..., d\}, \text{counting measure})$,
$\{p_\mu(\omega) = \mu_\omega, \mu \in \mathcal{M} = \{\mu > 0 : \sum_{\omega=1}^{d} \mu_\omega = 1\}$, $\mathcal{F} = \{\text{all functions on } \Omega\}$

Direct product of good o.s. $\mathcal{O}_t = ((\Omega_t, P_t), \{p_{\mu_t,t}(\cdot) : \mu_t \in \mathcal{M}_t\}, \mathcal{F}_t), 1 \leq t \leq K$

Samples ω^K of K independent observations drawn from $\mathcal{O}_1, ..., \mathcal{O}_K$:

$$(\Omega, P) = (\bigotimes_{t=1}^{K} \Omega_t, \bigotimes_{t=1}^{K} P_t), \{p_\mu(\omega^K) = \prod_{t=1}^{K} p_{\mu_t,t}(\omega_t) : \mu \in \mathcal{M} := \bigotimes_{t=1}^{K} \mathcal{M}_t\},$$

$$\mathcal{F}^{(K)} = \{f(\omega^K) = \sum_{t=1}^{K} f_t(\omega_t) : f_t \in \mathcal{F}_t\}$$

K-repeated version of a good o.s. $\mathcal{O} = ((\Omega, P), \{p_\mu(\cdot) : \mu \in \mathcal{M}\}, \mathcal{F})$

K-element i.i.d. samples ω^K drawn from $\mathcal{O}$:

$$(\Omega, P) = (\underbrace{\Omega \times ... \times \Omega}_{K}, \underbrace{P \times ... \times P}_{K}), \{p_\mu(\omega^K) = \prod_{t=1}^{K} p_\mu(\omega_t) : \mu \in \mathcal{M}\},$$

$$\mathcal{F}^{(K)} = \{\sum_{t=1}^{K} f(\omega_t) : f \in \mathcal{F}\}$$

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

♠ **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 $\mathcal{M}$.

- *Problem*

$$\ln(\epsilon_\star) = \max_{\mu\in X_1,\nu\in X_2}\ln\left(\int_\Omega\sqrt{p_\mu(\omega)p_\nu(\omega)}P(d\omega)\right)$$

is convex and solvable, and its optimal solution (μ_*, ν_*) induces the detector $\phi_*(\omega) = \frac{1}{2}\ln(p_{\mu_*}(\omega)/p_{\nu_*}(\omega))$ which is a balanced optimal solution to $(!)$.

- *For every K, the detector $\phi_*^{(K)}(\omega^K) = \sum_{t=1}^K \phi_*(\omega_t)$ induces test $\mathcal{T}^K$ deciding on the hypotheses $\mathbf{H}_1, \mathbf{H}_2$:*

$$\mathbf{H}_\chi : \omega_1, ..., \omega_K \text{ are i.i.d. drawn from } p_\mu \text{ with some } \mu \in X_\chi,$$

with risk $\epsilon_\star^K$, and this test is near-optimal: if "in the nature" there exists a test, based on K_* observations, deciding on $\mathbf{H}_1, \mathbf{H}_2$ with risk $\epsilon < 1/2$, the test $\mathcal{T}^K$ ensures the same risk ϵ whenever

$$K \geq \frac{2}{1-\frac{\ln(4(1-\epsilon))}{\ln(1/\epsilon)}}K_*$$

♠ **Note:** *For Discrete o.s.*, Main Theorem is covered by classical results of Le Cam, Huber & Strassen, and L. Birgé on deciding on two *convex* families of probability distributions.

♡ The novelty in the general case stems from the fact that *for a convex set X in the space of parameters of a good o.s., the associated family of distributions $\mathcal{P}_X = \{p_\mu(\cdot) : \mu \in X\}$ typically is nonconvex,* the Discrete o.s. being an exception.

From pairwise to multiple hypothesis testing

♣ **Recovery up to closeness:** *Given*

- *a good o.s.* $\mathcal{O} = ((\Omega, P), \{p_\mu(\cdot) : \mu \in \mathcal{M}\}, \mathcal{F})$
- *n convex compact sets* $X_i \subset \mathcal{M}$, $i = 1, ..., n$,
- *closeness* $\mathcal{C}$ *– symmetric Boolean* $n \times n$ *matrix with zero diagonal,*

along with an i.i.d. sample $\omega^K = (\omega_1, ..., \omega_K)$ *drawn from a distribution* p_{μ_*} *with some* $\mu_* \in \bigcup_i X_i$, *we want to decide on the hypotheses* $\mathbf{H}_i : \mu \in X_i$, $1 \leq i \leq n$, *"up to closeness* $\mathcal{C}$*"*, i.e., *we are ready to accept along with the true hypothesis* $\mathbf{H}_{i_*}$ (one with $\mu_* \in X_{i_*}$) *the hypotheses* $\mathbf{H}_i$ $\mathcal{C}$*-close to* $\mathbf{H}_{i_*}$ (those with $\mathcal{C}_{i_* i} = 0$).

♣ **Theorem.** *Tests* $\phi_{ij}(\cdot)$ *and risks* ϵ_{ij} *given by Main Theorem as applied to pairs of hypotheses* $\mathbf{H}_i$, $\mathbf{H}_j$, $1 \leq i < j \leq n$, *can be efficiently assembled into a test* $\mathcal{T}^K$ *deciding, up to closeness* $\mathcal{C}$, *on* $\mathbf{H}_1$,...,$\mathbf{H}_n$ *with risk at most*

$$\epsilon_\star = \left\| \left[\mathcal{C}_{ij} \epsilon_{ij}^K \right]_{i,j} \right\|_{2,2} \qquad [\epsilon_{ii} := 0, \epsilon_{ij} := \epsilon_{ji} \text{ for } i > j]$$

meaning that

As applied to i.i.d. sample $\omega_t \sim p_{\mu_*}(\cdot)$, $1 \leq t \leq K$, *with* $\mu_* \in X_{i_*}$ *for some* i_*, *the* p_{μ_*}*-probability of the event "*$\mathcal{T}^K$ *accepts the true hypothesis* $\mathbf{H}_{i_*}$, *and all hypotheses accepted by* $\mathcal{T}^K$ *are* $\mathcal{C}$*-close to* $\mathbf{H}_{i_*}$*" is at least* $1 - \epsilon_\star$.

♠ **Follow up:** *The test $\mathcal{T}^K$ is near-optimal:*

♠ *Assume that in the nature there exists a test, based on K_*-repeated observations, solving the Recovery-up-to-closeness problem with risk $\epsilon < 1/2$. Then the test $\mathcal{T}^K$ solves the same problem with the same risk ϵ whenever the number of observations K satisfies*

$$K \geq \frac{2\ln(n/\epsilon)}{\ln(1/\epsilon) - \ln(4(1-\epsilon))} K_*.$$

♠ Along with "static" versions of the above testing problems, we can address their

• *sequential settings*, where at time instants $t = 1, 2, ..., K$, given observations $\omega_1, ..., \omega_t$ acquired so far, we either make inference and terminate, or pass to the next observation, and the goal is to make reliable inference as fast as possible;

• *dynamical settings*, where the hypotheses "evolve in time" (change point detection)

♠ We can utilize tests to solve in a near-optimal, in the minimax sense, *estimation problems* like

> *Given a finite collection of convex compact sets $\mathcal{X}_i \subset \mathcal{M}$ and a function $f : \mathcal{X} := \bigcup_i \mathcal{X}_i \to \mathbb{R}$ which is affine (or affine-fractional) on every one of $\mathcal{X}_i$, estimate $f(\mu)$ from an i.i.d. sample $\omega_t \sim p_\mu$, $1 \leq t \leq K$, with unknown μ known to belong to $\mathcal{X}$.*

♣ **Potential applications:** design of near-optimal tests and estimates in

♠ **Gaussian Signal Processing**, where, given an observation

$$\omega = Ax + \mathcal{N}(0, I)$$

of unknown signal $x \in \bigcup_{i=1}^{N} \mathcal{X}_i$ with convex compact $\mathcal{X}_i$, we want to make inferences on the "location" of x

♠ **Poisson Imaging** – same as Gaussian Signal Processing, but with observation ω with independent entries $\omega_i \sim \text{Poisson}([Ax]_i)$, where $A \geq 0$ and $\mathcal{X}_i \subset \mathbb{R}^n_+$.

Poisson Imaging covers image recovery problems in

- 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.