On Statistical Inferences via Convex Optimization

A. Nemirovski
Georgia Institute of Technology

joint research with

Anatoli Juditsky
Université Grenoble Alpes

Workshop on the Interface of Statistics and Optimization
The Statistical and Applied Mathematical Sciences Institute
February 7, 2017

♣ **Fact:** Many inference procedures in Statistics reduce to optimization

♠ **Example: MLE – Maximum Likelihood Estimation**

> **Problem:** *Given a parametric family $\{p_\theta(\cdot) : \theta \in \Theta\}$ of probability densities on $\mathbb{R}^d$ and a random observation ω drawn from some density $p_{\theta_\star}(\cdot)$ from the family, estimate the parameter $\theta_\star$.*
>
> **Maximum Likelihood Estimate:** Given ω, maximize $p_\theta(\omega)$ over $\theta \in \Theta$ and use the maximizer $\widehat{\theta} = \widehat{\theta}(\omega)$ as an estimate of $\theta_\star$.

Note: In MLE, optimization is used for number crunching only and has nothing to do with motivation and performance analysis of MLE.

♣ Most of traditional applications of Optimization in Statistics are of "number crunching" nature.

• *In contrast, we will focus on inference routines motivated and justified by Optimization Theory* – Convex Analysis, Optimality Conditions, Duality...

Detector-Based Hypothesis Testing

Detectors & Detector-Based Pairwise Tests

♣ **Situation:** *Given two families $\mathcal{P}_1$, $\mathcal{P}_2$ of probability distributions on an 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).*

♣ **Detectors.** A *detector* is a function $\phi : \Omega \to \mathbb{R}$. *Risks* of a detector ϕ w.r.t. $\mathcal{P}_1, \mathcal{P}_2$ are defined as

$$\begin{array}{c}
\mathrm{Risk}_1(\phi|\mathcal{P}_1,\mathcal{P}_2) = \sup\limits_{P\in\mathcal{P}_1} \int_\Omega \mathrm{e}^{-\phi(\omega)} P(d\omega), \\
\mathrm{Risk}_2(\phi|\mathcal{P}_1,\mathcal{P}_2) = \sup\limits_{P\in\mathcal{P}_2} \int_\Omega \mathrm{e}^{\phi(\omega)} P(d\omega) \\
\mathrm{Risk}_1(\phi|\mathcal{P}_1,\mathcal{P}_2) = \mathrm{Risk}_2(-\phi|\mathcal{P}_2,\mathcal{P}_1)
\end{array}$$

♠ **Simple test** $\mathcal{T}_\phi$ associated with detector ϕ, given observation ω,

- accepts H_1 – $\mathcal{T}_\phi(\omega) = 1$ – when $\phi(\omega) \geq 0$,
- accepts H_2 – $\mathcal{T}_\phi(\omega) = 2$ – when $\phi(\omega) < 0$.

♣ **Immediate observation:**

$$\boxed{\begin{array}{rcl}
\mathrm{Risk}_1[\mathcal{T}_\phi|H_1,H_2] & \leq & \mathrm{Risk}_1(\phi|\mathcal{P}_1,\mathcal{P}_2) \\
\mathrm{Risk}_2[\mathcal{T}_\phi|H_1,H_2] & \leq & \mathrm{Risk}_2(\phi|\mathcal{P}_1,\mathcal{P}_2)
\end{array}} \qquad (*)$$

where test's risks Risk_1, Risk_2 are

$$\mathrm{Risk}_\chi[\mathcal{T}_\phi|H_1,H_2] = \sup_{P\in\mathcal{P}_\chi} \mathrm{Prob}_{\omega\sim P}\left\{\mathcal{T}_\phi(\omega) \neq \chi\right\}$$

Reason for $(*)$**:** $\mathrm{Prob}_{\omega\sim P}\{\omega : \psi(\omega) \geq 0\} \leq \int \mathrm{e}^{\psi(\omega)} P(d\omega)$.

$$\text{Risk}_1(\phi|\mathcal{P}_1,\mathcal{P}_2) = \sup_{P\in\mathcal{P}_1} \int_\Omega e^{-\phi(\omega)} P(d\omega),$$
$$\text{Risk}_2(\phi|\mathcal{P}_1,\mathcal{P}_2) = \sup_{P\in\mathcal{P}_2} \int_\Omega e^{\phi(\omega)} P(d\omega)$$

♣ Detectors admit simple "calculus:"

♣ **Renormalization:** $\phi(\cdot) \Rightarrow \phi_a(\cdot) = \phi(\cdot) - a$

$$\Rightarrow \left\{ \begin{array}{rcl} \text{Risk}_1(\phi_a|\mathcal{P}_1,\mathcal{P}_2) & = & e^a \text{Risk}_1(\phi|\mathcal{P}_1,\mathcal{P}_2) \\ \text{Risk}_2(\phi_a|\mathcal{P}_1,\mathcal{P}_2) & = & e^{-a} \text{Risk}_2(\phi|\mathcal{P}_1,\mathcal{P}_2) \end{array} \right.$$

$\Rightarrow$ *What matters, is the product*

$$[\text{Risk}(\phi|\mathcal{P}_1,\mathcal{P}_2)]^2 := \text{Risk}_1(\phi|\mathcal{P}_1,\mathcal{P}_2)\text{Risk}_2(\phi|\mathcal{P}_1,\mathcal{P}_2)$$

of partial risks of a detector. Shifting the detector by constant, we can distribute this product between factors as we want, e.g., always can make the detector balanced:

$$\text{Risk}(\phi|\mathcal{P}_1,\mathcal{P}_2) = \text{Risk}_1(\phi|\mathcal{P}_1,\mathcal{P}_2) = \text{Risk}_2(\phi|\mathcal{P}_1,\mathcal{P}_2).$$

♣ **Passing to multiple observations.** For $1 \le k \le K$, let

• $\mathcal{P}_{1,k}, \mathcal{P}_{2,k}$ be families of probability distributions on observation spaces Ω_k,

• ϕ_k be detectors on Ω_k.

♡ Families $\{\mathcal{P}_{1,k}, \mathcal{P}_{2,k}\}_{k=1}^K$ give rise to families of product distributions on $\Omega^K = \Omega_1 \times ... \times \Omega_K$:

$$\mathcal{P}_\chi^K = \{P^K = P_1 \times ... \times P_K : P_k \in \mathcal{P}_{\chi,k},\ 1 \le k \le K\},\ \chi = 1, 2,$$

and detectors $\phi_1, .., \phi_K$ give rise to detector ϕ^K on Ω^K:

$$\phi^K(\underbrace{\omega_1, ..., \omega_K}_{\omega^K}) = \sum_{k=1}^K \phi_k(\omega_k).$$

♠ **Observation:** *We have*

$$\mathrm{Risk}_\chi(\phi^K|\mathcal{P}_1^K, \mathcal{P}_2^K) = \prod_{k=1}^K \mathrm{Risk}_\chi(\phi_k|\mathcal{P}_{1,k}, \mathcal{P}_{2,k}).$$

♣ From pairwise detectors to detectors for unions

Assume that we are given an observation space Ω along with

- R families $\mathcal{R}_r$, $r = 1, ..., R$ of "red" probability distributions on Ω,
- B families $\mathcal{B}_b$, $b = 1, ..., B$ of "brown" probability distributions on Ω,
- pairwise detectors $\phi_{rb}(\cdot)$, $1 \leq r \leq R$, $1 \leq b \leq B$.

$$\epsilon_{rb} := \text{Risk}(\phi_{rb}|\mathcal{R}_r, \mathcal{B}_b) = \text{Risk}_1(\phi_{rb}|\mathcal{R}_r, \mathcal{B}_b) = \text{Risk}_2(\phi_{rb}|\mathcal{R}_r, \mathcal{B}_b),$$

Let us aggregate the red and the brown families as follows

$$\mathcal{R} = \bigcup_{r=1}^{R} \mathcal{R}_r, \ \mathcal{B} = \bigcup_{b=1}^{B} \mathcal{B}_b$$

and consider matrices

$$E = \left[\begin{array}{ccc} \epsilon_{1,1} & \cdots & \epsilon_{1,B} \\ \vdots & \cdots & \vdots \\ \epsilon_{R,1} & \cdots & \epsilon_{R,B} \end{array}\right], \ F = \left[\begin{array}{c|c} & E \\ \hline E^T & \end{array}\right]$$

The maximal eigenvalue of F is the spectral norm $\|E\|_{2,2}$ of E, and the leading eigenvector $[g; f]$ can be selected to be positive, giving rise to shifted detectors

$$\psi_{rb}(\omega) = \phi_{rb}(\omega) - \ln(f_b/g_r)$$

which can further be assembled into the detector

$$\psi(\omega) = \max_{r \leq R} \min_{b \leq B} \psi_{rb}(\omega)$$

Theorem: *Partial risks of detector ψ on aggregated families $\mathcal{R}$, $\mathcal{B}$ are $\leq \|E\|_{2,2}$.*

Detector-Based Tests ”Up to Closeness”

♣ **Situation:** We are given L families of probability distributions $\mathcal{P}_\ell$, $1 \leq \ell \leq L$, on observation space Ω, and observe a realization of random variable $\omega \sim P$ taking values in Ω. Given ω, we want to decide on the L hypotheses

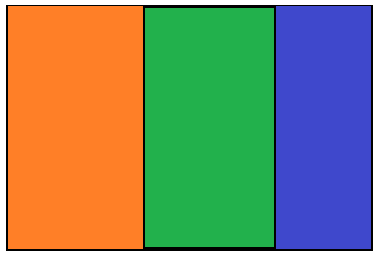
$$H_\ell : P \in \mathcal{P}_\ell,\ 1 \leq \ell \leq L.$$

Our ideal goal would be to find a low-risk simple test deciding on the hypotheses.

However: It may happen that the “ ideal goal” is not achievable, for example, when some pairs of families $\mathcal{P}_\ell$ have nonempty intersections. When $\mathcal{P}_\ell \cap \mathcal{P}_{\ell'} \neq \emptyset$ for some $\ell \neq \ell'$, there is no way to decide on the hypotheses with risk $< 1/2$.

But: *Impossibility to decide reliably on all L hypotheses “individually” does not mean that no meaningful inferences can be done.*

♠ **Example:** Consider 3 colored rectangles on the plane:



and 3 hypotheses, with H_ℓ, $1 \leq \ell \leq 3$, stating that our observation is $\omega = x + \xi$ with deterministic "signal" x belonging to ℓ-th rectangle and $\xi \sim \mathcal{N}(0, \sigma^2 I_2)$.
♡ Whatever small σ be, no test can decide on the 3 hypotheses with risk $< 1/2$; e.g., there is no way to decide reliably on H_1 vs. H_2.
However, *we may hope that when σ is small, we can discard reliably some of the hypotheses. For example, if the actual signal is brown, we cannot exclude the possibility for it to be claimed green, but hopefully can infer that it is not blue.*

♠ When handling multiple hypotheses which cannot be reliably decided upon "as they are," it makes sense to speak about *testing the hypotheses "up to closeness."*

♠ **Situation:** We are given

• L families of probability distributions $\mathcal{P}_\ell$, $\ell = 1, ..., L$, on observation space Ω, giving rise to L hypotheses H_ℓ, on the distribution P of random observation ω $in \Omega$:

$$H_\ell : P \in \mathcal{P}_\ell,\ 1 \leq \ell \leq L;$$

• *closeness relation $\mathcal{C}$ – a set $\mathcal{C}$ of pairs (ℓ, ℓ') of indexes of "close to each other" hypotheses H_ℓ, $H_{\ell'}$ such that $(\ell, \ell) \in \mathcal{C}$ (every hypothesis is close to itself) and $(\ell, \ell') \in \mathcal{C}$ whenever $(\ell', \ell) \in \mathcal{C}$ (closeness is symmetric).*

• system of balanced detectors

$$\left\{\phi_{\ell\ell'} : \ell < \ell', (\ell, \ell') \notin \mathcal{C}\right\}$$

along with upper bounds $\epsilon_{\ell\ell'}$ on detectors' risks:

$$\forall (\ell, \ell' : \ell < \ell', (\ell, \ell') \notin \mathcal{C}) : \left\{ \begin{array}{l} \int_\Omega \mathrm{e}^{-\phi_{\ell\ell'}(\omega)} P(d\omega) \leq \epsilon_{\ell\ell'}\ \forall P \in \mathcal{P}_\ell \\ \int_\Omega \mathrm{e}^{\phi_{\ell\ell'}(\omega)} P(d\omega) \leq \epsilon_{\ell\ell'}\ \forall P \in \mathcal{P}_{\ell'} \end{array}\right.$$

• Our goal is to build single-observation test deciding on hypotheses $H_1, ..., H_L$ up to closeness $\mathcal{C}$.

♠ **Definition.** *Let $\mathcal{T}$ be a test which, given observation ω, accepts some of the hypotheses H_ℓ and rejects the remaining hypotheses. We say that $\mathcal{C}$-risk of $\mathcal{T}$ is $\leq \epsilon$, if, whenever the distribution P of the observation obeys H_{ℓ_*} for some $\ell_* \leq L$, the P-probability of the event "H_{ℓ_*} is accepted, and all accepted hypotheses are $\mathcal{C}$-close to H_{ℓ_*}" is at least $1 - \epsilon$.*

♠ **Proposition.** *The pairwise detectors $\phi_{\ell\ell'}$ can be straightforwardly assembled into single-observation test $\mathcal{T}$ with $\mathcal{C}$-risk upper-bounded by*

$$\left\|[\epsilon_{\ell\ell'}\chi_{\ell\ell'}]_{\ell,\ell'=1}^{L}\right\|_{2,2} \qquad \left[\chi_{\ell\ell'} = \left\{\begin{array}{ll} 1, & (\ell,\ell')\not\in\mathcal{C} \\ 0, & (\ell,\ell')\in\mathcal{C}\end{array}\right.\right],$$

♠ **Corollary.** *Let $\epsilon_{\ell\ell'} \leq \theta < 1$ whenever $(\ell,\ell') \not\in \mathcal{C}$ and let stationary K-repeated observations – i.i.d. samples*

$$\omega^K = (\omega_1, ..., \omega_K)$$

drawn from distributions in question – be allowed. Then the K-repeated version $\mathcal{T}^K$ of $\mathcal{T}$ – with detectors

$$\phi_{\ell\ell'}^{(K)}(\omega^K) = \sum_{t=1}^{K} \phi_{\ell\ell'}(\omega_t)$$

in the role of $\phi_{\ell\ell'}$ – satisfies

$$\text{Risk}^{\mathcal{C}}[\mathcal{T}^K|H_1, ..., H_L] \leq \theta^K L.$$

♣ **"Universality" of detector-based tests.** *Let $\mathcal{P}_\chi$, $\chi = 1,2$, be two families of probability distributions on observation space Ω,and H_χ, $\chi = 1,2$, be associate hypotheses on the distribution of an observation.*

Assume that there exists a simple deterministic or randomized test $\mathcal{T}$ deciding on H_1, H_2 with risk $\leq \epsilon \in (0,1/2)$. Then there exists a detector ϕ with

$$\text{Risk}(\phi|\mathcal{P}_1,\mathcal{P}_2) \leq \epsilon_+ := 2\sqrt{\epsilon[1-\epsilon]} < 1.$$

♠ **Note:** Risk $2\sqrt{\epsilon[1-\epsilon]}$ of the detector-based test induced by simple test $\mathcal{T}$ is "much worse" than the risk ϵ of $\mathcal{T}$.

However: *When repeated observations are allowed, we can compensate for risk deterioration $\epsilon \mapsto 2\sqrt{\epsilon[1-\epsilon]}$ by passing in the detector-based test from a single observation to a moderate number of them.*

$$\inf_{\phi} \mathrm{Risk}(\phi|\mathcal{P}_1,\mathcal{P}_2) = \min\left\{\epsilon : \begin{array}{rcl} \int_\Omega e^{-\phi(\omega)}P(d\omega) & \leq & \epsilon\,\forall(P\in\mathcal{P}_1) \\ \int_\Omega e^{\phi(\omega)}P(d\omega) & \leq & \epsilon\,\forall(P\in\mathcal{P}_2) \end{array}\right\} \qquad (!)$$

Note:

- The optimization problem specifying risk is *convex* in ϕ, ϵ
- When passing from families $\mathcal{P}_\chi$, $\chi = 1,2$, to their convex hulls, the risk of a detector remains intact.

♣ **Intermediate conclusion:** *It would be nice to solve* (!)*, thus arriving at the lowest risk detector-based tests.*

But: (!) is an optimization problem with *infinite-dimensional* decision "vector" and *infinitely many* constraints.

⇒ (!) *in general is intractable.*

Simple observation schemes: A series of special cases where (!) is efficiently solvable via Convex Optimization.

Simple Observation Schemes

♣ **Simple Observation Scheme** admits a formal definition which we skip.
Instructive examples are as follows.

♠ **Gaussian o.s.**:

$$\omega = A(x) + \mathcal{N}(0, I_d)$$

• $A(x)$: affine image of unknown *signal* x varying in *signal space* $\mathcal{X} := \mathbb{R}^n$.
• Gaussian o.s. is the standard observation model in Signal Processing.

♠ **Poisson o.s.**:

$\omega \in \mathbf{Z}^d$, $\omega_i \sim \text{Poisson}[A_i(x)]$ independent across $i = 1, ..., d$

• $A_i(x)$: affine functions of unknown *signal* x varying in a given open convex *signal space* $\mathcal{X} \subset \mathbb{R}^n$ such that $A_i(x) > 0$, $x \in \mathcal{X}$.

Poisson o.s. arises in *Poisson Imaging*, including

- *Positron Emission Tomography*,
- *Large binocular Telescope*,
- *Nanoscale Fluorescent Microscopy*.

♠ **Discrete o.s.**:

$$\omega \in \{e_1, ..., e_d\} \text{ takes value } e_i \text{ with probability } A_i(x)$$

• e_i: i-th basic orth in $\mathbb{R}^d$

• A_i: affine functions of unknown *signal* x varying in a given open convex *signal space* $\mathcal{X} \subset \mathbb{R}^n$ such that $A_i(x) > 0$ and $\sum_i A_i(x) = 1$, $x \in \mathcal{X}$.

♠ K**-repeated version** of a simple o.s.:

$$\omega = \Omega^K := (\omega_1, ..., \omega_K)$$

with ω_t sampled, independently across t, from observations of an unknown signal $x \in \mathcal{X}$ yielded by a simple o.s., e.g., Gaussian/Poisson/Discrete one.

♠ **Note:** Distributions P of observations in a simple o.s. possess positive continuous densities $p(\cdot)$ w.r.t. a properly selected reference measure Π on the space of observations.

♠ **Convex hypothesis** H_X in a simple o.s. is specified by a *nonempty convex compact* subset X of the corresponding signal space $\mathcal{X}$ and states that the signal x underlying observation belongs to X.

$$\varepsilon_\star(\mathcal{P}_1,\mathcal{P}_2) = \min\left\{\epsilon : \begin{array}{lcl} \int_\Omega e^{-\phi(\omega)} P(d\omega) & \leq & \epsilon \,\forall (P\in\mathcal{P}_1) \\ \int_\Omega e^{\phi(\omega)} P(d\omega) & \leq & \epsilon \,\forall (P\in\mathcal{P}_2) \end{array}\right\} \qquad (!)$$

♣ **Main Result.** *For $\chi = 1,2$, let $\mathcal{P}_\chi$ of probability distributions obeying convex hypothesis $H_\chi : x \in X_\chi$ in a simple o.s. The problem*

$$\begin{array}{rcl} \mathrm{Opt} & = & \max\limits_{p_1,p_2}\left\{\int\sqrt{p_1(\omega)p_2(\omega)}\Pi(d\omega) : p_\chi(\cdot) \text{ is the density}\right. \\ & & \left.\text{of a distribution from } \mathcal{P}_\chi,\ \chi=1,2\right\} \end{array} \qquad (!)$$

is equivalent to an explicit finite-dimensional convex program and is solvable. Optimal solution $(p_1^*(\cdot), p_2^*(\cdot))$ to the problem gives rise to the minimum risk balanced detector

$$\phi_*(\omega) = \frac{1}{2}\ln(p_1^*(\omega)/p_2^*(\omega))$$

for $\mathcal{P}_1$, $\mathcal{P}_2$. This detector is an affine function of ω, and the risk of the detector is Opt.

• *In our standard o.s.'s,* $(!)$ *reads:*

• Gaussian o.s.:	$\ln(\mathrm{Opt}) = -\frac{1}{8}\min\limits_{x\in X_1, y\in X_2}\|A(x)-A(y)\|_2^2$ [Π: Lebesque measure]
• Poisson o.s.:	$\ln(\mathrm{Opt}) = -\frac{1}{2}\min\limits_{x\in X_1, y\in X_2}\sum_i [A_i^{1/2}(x) - A_i^{1/2}(y)]^2$ [Π: counting measure]
• Discrete o.s.:	$\mathrm{Opt} = \max\limits_{x\in X_1, y\in X_2}\sum_i A_i^{1/2}(x)A_i^{1/2}(y)$ [Π: counting measure]

For K-repeated version of a simple o.s., the optimal detector is

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

and its risk is $\mathrm{Opt}_K = \mathrm{Opt}^K$.

Near-Optimality of Minimum Risk Detector-Based Tests in Simple Observation Schemes

♣ **Proposition A.** *Let* H_{X_χ}*,* $\chi = 1, 2$*, be convex hypotheses in a simple o.s., and* $\mathcal{P}_\chi$ *be the family of distributions obeying the hypotheses. Assume that in the nature there exists a simple single-observation test* $\mathcal{T}$*, deterministic or randomized,* $\mathcal{T}$ *with*

$$\text{Risk}[\mathcal{T}|H_1, H_2] \leq \epsilon < 1/2.$$

Then the risk of the simple test $\mathcal{T}_{\phi_*}$ *accepting* H_1 *when* $\phi_*(\omega) \geq 0$ *and accepting* H_2 *otherwise is comparable to* ϵ*:*

$$\text{Risk}[\mathcal{T}_{\phi_*}|H_1, H_2] \leq \epsilon_+ := 2\sqrt{\epsilon(1-\epsilon)} < 1.$$

♣ **Proposition B.** *Let H_χ, $\chi = 1,2$, be convex hypotheses in a simple o.s. Assume that for some $\epsilon < 1/2$ and K_* in the nature there exists a test, based on K_*-repeated observations, with risk $\leq \epsilon$. Then the risk of the test $\mathcal{T}_{\phi_*^{(K)}}$ with*

$$K \geq \widehat{K}_* = 2 \underbrace{\left[\frac{\ln(1/\epsilon)}{\ln(1/\epsilon) - \ln(4(1-\epsilon))}\right]}_{\to 1 \text{ as } \epsilon \to +0} K_*.$$

does not exceed ϵ as well.

♣ **Proposition C.** *Let H_ℓ, $\ell = 1, 2, ..., L$, be convex hypotheses in a simple o.s., and $\mathcal{C}$ be a closeness relation. Assume that for some $\epsilon < 1/2$ and K_* in the nature there exists a test, based on K_*-repeated observations, deciding on the hypotheses with $\mathcal{C}$-risk $\leq \epsilon$. Then the efficiently computable K-observation test $\mathcal{T}^K$ yielded by assembling optimal pairwise detectors with*

$$K \geq 2 \underbrace{\left[\frac{\ln(1/\epsilon) + \ln(L-1)}{\ln(1/\epsilon) - \ln(4(1-\epsilon))}\right]}_{\to 1 \text{ as } \epsilon \to +0} K_*.$$

has $\mathcal{C}$-risk $\leq \epsilon$ as well.

♣ **Generic applications** of minimum-risk-detector-based tests in simple o.s. include

- near-optimal estimation of linear/factional-linear functionals on finite unions of convex signal sets
- sequential testing of multiple convex hypotheses
- change point detection in linear dynamical systems
- rudimentary measurement design

Illustration: Estimating Fractional-Linear Functional on Union of Convex Sets

♠ **Situation:** *Signal* x *known to belong to the finite union*

$$X = \bigcup_{\mu=1}^{M} X_\mu$$

of given convex compact sets X_μ *is observed via a Simple o.s. Given a linear-fractional function*

$$F(u) = \frac{a^T u + b}{c^T u + d} : X \to \mathbb{R}, \qquad \left[\min_{u \in X} c^T u + d > 0\right]$$

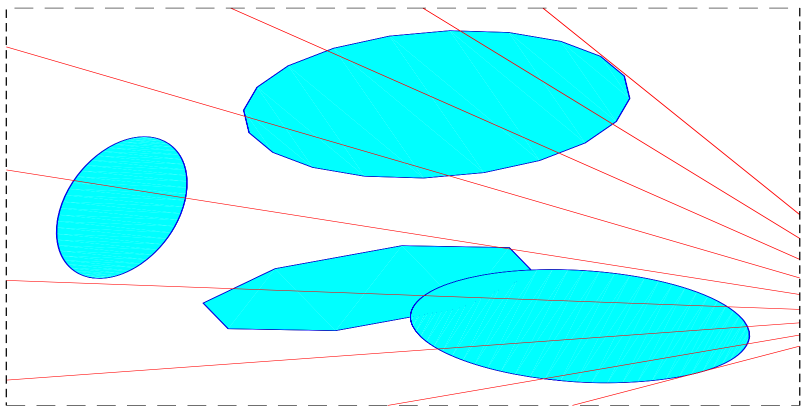
we want to recover $f(x)$ *via observation(s) associated with* x.

♠ **Strategy:** Given N, we

• split the range $\Delta = [\min_{x\in X} F(x), \max_{x\in X} F(x)]$ into N consecutive bins Δ_ν of length $\delta_N = |\Delta|/N$,

• define MN convex hypotheses

$$H_{\mu\nu} : x \in X_\mu \ \&\ F(x) \in \Delta_\nu$$



• use pairwise optimal detectors to decide on the convex hypotheses $H_{\mu\nu}$, $1 \leq \mu \leq M$, $1 \leq \nu \leq N$ up to closeness

$$\mathcal{C} : H_{\mu\nu} \text{ is close to } H_{\mu'\nu'} \Leftrightarrow \Delta_\nu \cap \Delta_{\nu'} \neq \emptyset$$

• estimate $F(x)$ by the center of masses $\widehat{F}$ of the union of bins Δ_ν associated with the accepted hypotheses $H_{\mu\nu}$.

♠ **Fact:** *For the resulting test $\mathcal{T}$ the recovery error does not exceed δ_N with probability at least $1-\mathrm{Risk}^{\mathcal{C}}[\mathcal{T}|H_{1,1},...,H_{M,N}]$.*

♠ **Near-Optimality:** *Let $\epsilon \in (0,1/2)$. Assume in the nature there exists an estimator recovering $F(x)$, $x \in X$, $(1-\epsilon)$-reliably within accuracy $\delta_N/2$ via K_* observations. Then $\mathrm{Prob}\{|\widehat{F} - F(x)| > \delta_N\} \leq \epsilon$, provided that the number K of observations underlying $\widehat{F}$ satisfies*

$$K \geq 2\left[\frac{\ln(MN/\epsilon)}{\ln(1/\epsilon) - \ln(4(1-\epsilon))}\right] K_*.$$

♣ **Observation:** *A "common denominator" of minimum risk detectors for simple o.s.'s is their affinity in observations.*

♠ **Fact:** *Presumably good affine detectors can be found, in a computationally efficient way, in many important situations which are beyond simple o.s.'s.*

Setup

♣ Given an observation space $\Omega = \mathbb{R}^d$, consider a triple $\mathcal{H}, \mathcal{M}, \Phi$, where

• $\mathcal{H}$ is a nonempty closed convex set in Ω symmetric w.r.t. the origin,

• $\mathcal{M}$ is a closed convex set in some $\mathbb{R}^n$,

• $\Phi(h; \mu) : \mathcal{H} \times \mathcal{M} \to \mathbb{R}$ is a continuous function *convex in* $h \in \mathcal{H}$ and *concave in* $\mu \in \mathcal{M}$.

♣ $\mathcal{H}, \mathcal{M}, \Phi$ specify a family $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$ of probability distributions on Ω. A probability distribution P belongs to the family iff there exists $\mu \in \mathcal{M}$ such that

$$\ln\left(\int_\Omega e^{h^T\omega} P(d\omega)\right) \leq \Phi(h; \mu) \ \forall h \in \mathcal{H} \qquad (*)$$

We refer to μ ensuring $(*)$ as to *parameter* of distribution P.

• **Warning:** A distribution P may have many different parameters!

♡ We refer to triple $\mathcal{H}, \mathcal{M}, \Phi$ satisfying the above requirements as to *regular data*, and to $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$ – as to the *simple family of distributions* induced by these data.

♠ **Example 1: Gaussian and sub-Gaussian distributions.** When $\mathcal{M} = \{(u,\Theta)\} \subset \mathbb{R}^d \times \text{int}\,\mathbf{S}^d_+$ is a convex compact set such that $\Theta \succ 0$ for all $(u,\Theta) \in \mathcal{M}$, $\mathcal{H} = \mathbb{R}^d$ and $\Phi(h; u,\Theta) = h^T u + \frac{1}{2}h^T\Theta h$, $\mathcal{S} = \mathcal{S}[\mathcal{H},\mathcal{M},\Phi]$ contains all probability distributions P which are *sub-Gaussian with parameters* (u,Θ), meaning that

$$\ln\left(\int_\Omega e^{h^T\omega}P(d\omega)\right) \leq h^T u + \frac{1}{2}h^T\Theta h \;\forall h, \qquad (1)$$

and, in addition, the "parameter" (u,Θ) belongs to $\mathcal{M}$.

Note: Whenever P is sub-Gaussian with parameters (u,Θ), u is the expectation of P.

Note: $\mathcal{N}(u,\Theta) \in \mathcal{S}$ whenever $(u,\Theta) \in \mathcal{M}$; for $P = \mathcal{N}(u,\Theta)$, (1) is an identity.

♠ **Example 2: Poisson distributions.** When $\mathcal{M} \subset \mathbb{R}^d_+$ is a convex compact set, $\mathcal{H} = \mathbb{R}^d$ and

$$\Phi(h;\mu) = \sum_{i=1}^{d} \mu_i(e^{h_i} - 1),$$

$\mathcal{S} = \mathcal{S}[\mathcal{H},\mathcal{M},\Phi]$ contains distributions of all d-dimensional random vectors ω_i with independent across i entries $\omega_i \sim \text{Poisson}(\mu_i)$ such that $\mu = [\mu_1; ...; \mu_d] \in \mathcal{M}$.

♠ **Example 3: Discrete distributions.** When

$$\mathcal{M} = \{\mu \in \mathbb{R}^d : \mu \geq 0, \sum_j \mu_j = 1\}$$

is the probabilistic simplex in $\mathbb{R}^d$, $\mathcal{H} = \mathbb{R}^d$ and

$$\Phi(h;\mu) = \ln\left(\sum_{i=1}^{d} \mu_i e^{h_i}\right),$$

$\mathcal{S} = \mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$ contains all discrete distributions supported on the vertices of the probabilistic simplex.

♠ **Example 4: Distributions with bounded support.** Let $X \subset \mathbb{R}^d$ be a nonempty convex compact set with support function $\phi_X(\cdot)$:

$$\phi_X(y) = \max_{x \in X} y^T x : \mathbb{R}^d \to \mathbb{R}^d.$$

When $\mathcal{M} = X$, $\mathcal{H} = \mathbb{R}^d$ and

$$\Phi(h;\mu) = h^T\mu + \frac{1}{8}[\phi_X(h) + \phi_X(-h)]^2, \qquad (2)$$

$\mathcal{S} = \mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$ contains all probability distributions supported on X, and for such a distribution P, $\mu = \int_X \omega P(d\omega)$ is a parameter of P.

• **Note:** Conclusion in Example IV remains valid when function (2) is replaced with the smaller function

$$\Phi_G(h;\mu) = \min_{g \in G} \left[\mu^T(h-g) + \tfrac{1}{8}[\phi_X(h-g) + \phi_X(g-h)]^2 + \phi_X(g)\right].$$

[$G \ni 0$: convex compact set]

♣ **Main observation:** *When deciding on simple families of distributions, affine tests and their risks can be efficiently computed via Convex Programming:*

♡ **Theorem.** *Let $\mathcal{H}_\chi, \mathcal{M}_\chi, \Phi_\chi$, $\chi = 1, 2$, be two collections of regular data with common $\mathcal{H}_1 = \mathcal{H}_2 =: \mathcal{H}$, and let*

$$\Psi(h) = \max_{\mu_1 \in \mathcal{M}_1, \mu_2 \in \mathcal{M}_2} \underbrace{\frac{1}{2}[\Phi_1(-h; \mu_1) + \Phi_2(h, \mu_2)]}_{\Phi(h;\mu_1,\mu_2)} : \mathcal{H} \to \mathbb{R}$$

Then Ψ is efficiently computable continuous convex function, and for every $h \in \mathcal{H}$, setting

$$\phi(\omega) = h^T\omega + \underbrace{\frac{1}{2}\left[\max_{\mu_1 \in \mathcal{M}_1} \Phi_1(-h; \mu_1) - \max_{\mu_2 \in \mathcal{M}_2} \Phi_1(h; \mu_2)\right]}_{\varkappa},$$

one has

$$\text{Risk}(\phi|\mathcal{P}_1, \mathcal{P}_2) \leq \exp\{\Psi(h)\} \qquad [\mathcal{P}_\chi = \mathcal{S}[\mathcal{H}, \mathcal{M}_\chi, \Phi_\chi]]$$

In particular, if convex-concave function $\Phi(h; \mu_1, \mu_2)$ possesses a saddle point $h_*, (\mu_1^*, \mu_2^*)$ on $\mathcal{H} \times (\mathcal{M}_1 \times \mathcal{M}_2)$, the affine detector

$$\phi_*(\omega) = h_*^T\omega + \frac{1}{2}[\Phi_1(-h; \mu_1^*) - \Phi_2(h^*; \mu_2^*)]$$

admits risk bound

$$\text{Risk}(\phi|\mathcal{P}_1, \mathcal{P}_2) \leq \exp\{\Phi(h^*; \mu_1^*, \mu_2)\}$$

♣ **Example: Sub-Gaussian Direct Product case.** For $\chi = 1, 2$, let $U_\chi \subset \Omega = \mathbb{R}^d$ and $\mathcal{V}_\chi \subset \mathrm{int}\, \mathbf{S}^d_+$ be convex compact sets. Setting

$$\mathcal{M}_\chi = U_\chi \times \mathcal{V}_\chi,\ \ \Phi(h; u, \Theta) = h^T u + \frac{1}{2} h^T \Theta h : \mathcal{H} \times \mathcal{M}_\chi \to \mathbb{R},$$

the regular data $\mathcal{H} = \mathbb{R}^d, \mathcal{M}_\chi, \Phi$ specify the families

$$\mathcal{P}_\chi = \mathcal{S}[\mathbb{R}^d, U_\chi \times \mathcal{V}_\chi, \Phi]$$

of sub-Gaussian distributions with parameters from $U_\chi \times \mathcal{V}_\chi$.

♠ Saddle point problem responsible for the design of affine detector for $\mathcal{P}_1, \mathcal{P}_2$ reads

$$\mathrm{SadVal} = \min_{h \in \mathbb{R}^d} \max_{\substack{u_1 \in U_1, u_2 \in U_2 \\ \Theta_1 \in \mathcal{V}_1, \Theta_2 \in \mathcal{V}_2}} \frac{1}{2} \left[h^T (u_2 - u_1) + \frac{1}{2} h^T [\Theta_1 + \Theta_2] h \right]$$

The problem is efficiently solvable, and its solution yields affine detector ϕ_* with risk

$$\mathrm{Risk}(\phi_* | \mathcal{P}_1, \mathcal{P}_2) \le \exp\{\mathrm{SadVal}\}.$$

♡ **Note:** In the *symmetric case $\mathcal{V}_1 = \mathcal{V}_2$ the affine detector we end up with is the minimum risk detector for $\mathcal{P}_1, \mathcal{P}_2$.*

♠ **Beyond Direct Product case:** Let

$$\mathcal{Q}_\chi = \{(\mu, \Theta) \in \mathbb{R}^d \times \mathbf{S}^d_{++}\}, \chi = 1, 2$$
$$\left[\mathbf{S}^d_{++} = \{\Theta \in \mathbf{S}^d : \Theta \succ 0\}\right]$$

be convex compact sets. Applying Theorem, we can test the hypotheses

$$H_\chi : \omega \sim \mathcal{N}(\mu, \Theta) \text{ with } (\mu, \Theta) \in \mathcal{Q}_\chi, \ \chi = 1, 2$$

via affine detector readily given by the solution to an explicit convex-concave saddle point problem.

Note: *Utilizing sets $\mathcal{Q}_\chi$, we extend Gaussian o.s. by allowing for dependencies between the mean and the covariance of observations.*

What is "affine?" Quadratic Lifting

♣ We have developed a technique for building reasonable *affine* detectors for simple families of distributions.

But: Given observation $\zeta \sim P$, we can subject it to *nonlinear* transformation $\zeta \mapsto \omega = \psi(\zeta)$, e.g., *quadratic lifting*

$$\zeta \mapsto \omega = (\zeta, \zeta\zeta^T)$$

and treat as our observation ω rather than the "true" observation ζ. *Affine in ω detectors are nonlinear in ζ.*

Example: Detectors affine in the quadratic lifting $\omega = (\zeta, \zeta\zeta^T)$ of ζ are exactly the *quadratic* functions of ζ.

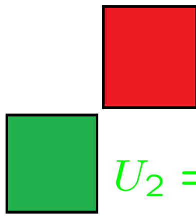
♠ We can try to apply our machinery for building affine detectors to nonlinear transformations of true observations, thus arriving at nonlinear detectors.

• **Bottleneck:** To apply the outlined strategy to a pair $\mathcal{P}_1, \mathcal{P}_2$ of families of distributions of interest, we need to cover the families $\mathcal{P}_\chi^+$ of distributions of $\omega = \psi(\zeta)$ induced by distributions $P \in \mathcal{P}_\chi$, $\chi = 1, 2$, by simple families of distributions.

♠ *The bottleneck can be resolved reasonably well for Gaussian and sub-Gaussian distributions.*

♡ Numerical illustration: Gaussian Direct Product case

$$\zeta = Au + \sigma\xi,\ \xi \sim \mathcal{N}(0, I_8)$$
$$H_\chi : u \in U_\chi \ \&\ \sigma = \sigma_\chi,\ \chi = 1, 2$$



$U_1 = U_1^\rho = \{u \in \mathbb{R}^{12} : u_i \geq \rho, 1 \leq i \leq 12\}$

$U_2 = U_2^\rho = -U_1^\rho$

• $A \in \mathbb{R}^{8\times 12}$ (*deficient observations*)

ρ	σ_1	σ_2	unrestricted H and h	$H = 0$	$h = 0$
0.5	2	2	0.31	0.31	1.00
0.5	1	4	0.24	0.39	0.62
0.01	1	4	0.41	1.00	0.41

Risk of quadratic detector $\phi(\zeta) = h^T\zeta + \frac{1}{2}\zeta^T H\zeta + \varkappa$

♣ We see that

• when deciding on families of Gaussian distributions with common covariance matrix and expectations varying in associated with the families convex sets, passing from affine to quadratic detectors does not help.

• in general, both affine and purely quadratic components in a quadratic detector are useful.

• when deciding on families of Gaussian distributions in the case where distributions from different families can have close expectations, affine detectors are useless, while the quadratic ones are not.

♣ **Model:** *We observe one by one vectors ("vectorized" 2D images)*

$$\omega_t = x_t + \xi_t,$$

- x_t: deterministic image
- $\xi_t \sim \mathcal{N}(0, \sigma^2 I_d)$: independent across observation noises.
  Note: We know a range $[\underline{\sigma}, \overline{\sigma}]$ of σ, but perhaps do not know σ exactly.

- We know that $x_1 = x_2$ and want to check whether $x_1 = ... = x_K$ ("no change") or there is a change.

♠ **Goal:** *Given an upper bound* $\epsilon > 0$ *on the probability of false alarm, we want to design a sequential change detection routine capable to detect change, if any.*

♠ **Approach:**

- Pass from observations ω_t, $1 \leq t \leq K$, to observations

$$\zeta_t = \omega_t - \omega_1 = \underbrace{x_t - x_1}_{y_t} + \underbrace{\xi_t - \xi_1}_{\eta_t},\ 2 \leq t \leq K$$

- Test null hypothesis H_0 "no change" $(y_2 = ... = y_K = 0)$ vs. alternative $\bigcup_{k=2}^{K} \{H_k^\rho$: change at time k of magnitude $\geq \rho\}$

$$H_k^\rho : y_2 = ... = y_{k-1} = 0, \|y_k\|_2 \geq \rho$$

via our machinery for testing multiple hypotheses $\mathcal{G}_k^\rho$ on quadratic lifts $Y_k = y_k y_k^T$ of observations y_k:

$$\begin{array}{c} \mathcal{G}_1 : \{Y_1 = ... = Y_K = 0\}, \\ \mathcal{G}_k^\rho : \{Y_1 = ... = Y_{k-1} = 0, \mathrm{Tr}(Y_k) \geq \rho^2, Y_t \succeq 0\ \forall t\},\ 2 \leq k \leq K \end{array}$$

up to closeness

> $\mathcal{C}$: *all brown hypotheses are close to each other and are not close to the magenta hypothesis*

- Find the smallest ρ for which the $\mathcal{C}$-risk of the resulting inference is $\leq \epsilon$, and utilize this inference in change point detection.

How It Works

♠ **Setup:** $\dim y = 256^2 = 65536$, $\overline{\sigma} = 10$, $\overline{\sigma}^2/\underline{\sigma}^2 = 2$, $K = 9$, $\epsilon = 0.01$

♠ **Inference:** At time $t = 2, ..., K$, compute

$$\phi_*(\zeta_t) = -2.7138\frac{\|\zeta_t\|_2^2}{10^5} + 366.9548.$$

$\phi_*(\zeta_t) < 0 \Rightarrow$	*conclude that the change took place and terminate*
$\phi_*(\zeta_t) \geq 0 \Rightarrow$	*conclude that there was no change so far and proceed to the next image, if any*

♠ **Note:**

- *When $\mathcal{G}_1$ holds true, the probability not to claim change on time horizon $1, ..., K$ is at least 0.99.*
- *When G_k^ρ holds true, the change at time $\leq k$ is detected with probability at least 0.99*, provided $\rho \geq \rho_* = 2716.6$ (average per pixel energy in y_k at least by 12% larger than $\overline{\sigma}^2$)
- *No test can 0.99-reliably decide via $\zeta_1, ..., \zeta_k$ on $\mathcal{G}_k^\rho$ vs. $\mathcal{G}_1$*, provided $\rho/\rho_* < 0.965$.
- *In the movie, the change takes place at time 3 and is detected at time 4.*

Signal Estimation in Gaussian O.S.

♣ **Situation:** "In the nature" there exists a signal x known to belong to a given convex compact set $\mathcal{X} \subset \mathbb{R}^n$. We observe corrupted by noise affine image of the signal:

$$\omega = Ax + \sigma\xi \in \mathbb{R}^m$$

•A: given $m \times n$ sensing matrix
•ξ: random observation noise

♠ **Goal:** To recover the image Bx of x under a given linear mapping

•B: given $k \times n$ matrix.

♠ **Risk** of a candidate estimate $\widehat{x}(\cdot) : \Omega \to \mathbb{R}^k$ is defined as

$$\mathrm{Risk}[\widehat{x}|\mathcal{X}] = \sup_{x \in \mathcal{X}} \sqrt{\mathbf{E}_\xi \left\{ \|Bx - \widehat{x}(Ax + \sigma\xi)\|_2^2 \right\}}$$

$\Rightarrow$ Risk^2 is the worst-case, over $x \in \mathcal{X}$, expected $\|\cdot\|_2^2$ recovery error.

♣ **Agenda:** Under appropriate assumptions on $\mathcal{X}$, we shall show that

• *One can build, in a computationally efficient fashion, the (nearly) best, in terms of risk, in the family of linear estimates*

$$\widehat{x}(\omega) = \widehat{x}_H(\omega) = H^T\omega \qquad [H \in \mathbb{R}^{m \times k}]$$

• *The resulting linear estimate is nearly optimal among all estimates, linear and nonlinear alike.*

♣ **Assumption on noise:** ξ *is zero mean with unit covariance matrix.*

$\Rightarrow$ *The risk of a linear estimate* $\widehat{x}_H(\omega) = H^T\omega$ *is given by*

$$\mathrm{Risk}^2[\widehat{x}_H|\mathcal{X}] = \sigma^2\mathrm{Tr}(H^TH) + \underbrace{\max_{x\in\mathcal{X}} \mathrm{Tr}([B - H^TA]xx^T[B^T - A^TH])}_{\Psi(H)}.$$

♡ **Note:** Ψ is convex $\Rightarrow$ *building the minimum risk linear estimate reduces to solving* convex *minimization problem*

$$\mathrm{Opt}_* = \min_H \left[\Psi(H) + \sigma^2\mathrm{Tr}(H^TH)\right]. \qquad (*)$$

But: Convex function Ψ is given implicitly and can be difficult to compute, making $(*)$ difficult as well.

Fact: Basically, the only cases when $(*)$ is known to be easy are those when

- $\mathcal{X}$ is given as a convex hull of finite set of moderate cardinality
- $\mathcal{X}$ is an ellipsoid.

$\mathcal{X}$ is a box $\Rightarrow$ computing Ψ is NP-hard...

$$\min_H \left\{ \sigma^2 \mathrm{Tr}(H^T H) + \underbrace{\max_{x\in\mathcal{X}} \mathrm{Tr}([B - H^T A]xx^T[B^T - A^T H])}_{\Psi(H)} \right\} \quad (*)$$

♠ When Ψ is difficult to compute, we can to replace Ψ in the design problem $(*)$ with an efficiently computable convex upper bound $\Psi^+(H)$.
We are about to consider a family of sets $\mathcal{X}$ – *ellitopes* – for which reasonably tight bounds Ψ^+ are available.

♣ **An ellitope** is a *bounded* set $\mathcal{X} \subset \mathbb{R}^n$ given as

$$\mathcal{X} = \{x \in \mathbb{R}^n : \exists y \in \mathbb{R}^N, t \in \mathcal{T} : x = Ry,\ y^T S_k y \preceq t_k,\ 1 \le k \le K\}$$

where

- R is a given $n \times N$ matrix,
- S_k are *positive semidefinite* matrices
- $\mathcal{T}$ is a convex compact subset of $\mathbb{R}^K_+$ containing a positive vector and *monotone*:

$$0 \le t' \le t \in \mathcal{T} \Rightarrow t' \in \mathcal{T}.$$

♠ **Note:** *Every ellitope is a symmetric w.r.t. the origin convex compact set.*

♠ **Basic examples:**

A. Intersection $\bigcap_i \{x : \|A_i x\|_2 \leq 1\}$ of finitely many ellipsoids/elliptic cylinders centered at the origin

B. Intersection $\bigcap_i \{x : \|A_i x\|_{p_i} \leq 1$ of finitely many "ℓ_p balls/cylinders" centered at the origin, with $2 \leq p_i \leq \infty$

♣ **Note:** *What follows straightforwardly extends from ellitopes to their "matrix analogies" – spectratopes*

$$\mathcal{X} = \{x \in \mathbb{R}^n : \exists (y \in \mathbb{R}^N, t \in \mathcal{T}) : x = Ry, S_k^2[y] \preceq t_k I_{d_k}, k \leq K\}$$
$[S_k[y]$: symmetric $d_k \times d_k$ matrices linearly depending on $y]$

Every ellitope is a spectratope, but not vice versa; e.g., the matrix box $\{y \in \mathbb{R}^{m \times n} : \text{ spectral norm of } y \text{ is } \leq 1\}$ *is a spectratope, but not an ellitope.*

♠ *Ellitopes/spectratopes admit fully algorithmic calculus:*

if $\mathcal{X}_i$, $1 \leq i \leq I$, are ellitopes/spectratopes, so are

- $\bigcap_i \mathcal{X}_i$
- $\mathcal{X}_1 \times ... \times \mathcal{X}_I$
- $\mathrm{Conv}(\bigcup_i \mathcal{X}_i)$
- $\mathcal{X}_1 + ... + \mathcal{X}_I$
- linear images of $\mathcal{X}_i$
- inverse linear images of $\mathcal{X}_i$ under linear embeddings

♣ **Observation:** *It is easy to upper-bound the maximum of a quadratic form* x^TQx *over an ellitope*

$$\mathcal{X} = \{x : \exists (t \in \mathcal{T}, y) : x = Ry, y^T S_k y \leq t_k, 1 \leq k \leq K\}.$$

Specifically, whenever $\lambda \geq 0$ *satisfies*

$$R^T Q R \preceq \sum_k \lambda_k S_k,$$

we have

$$\max_{x \in \mathcal{X}} x^T Q x \leq \phi_{\mathcal{T}}(\lambda) := \max_{t \in \mathcal{T}} \lambda^T t.$$

♠ **Corollary:** *Given an ellitope $\mathcal{X}$ and matrices A, B, consider the convex optimization problem*

$$\mathrm{Opt} = \min_{\lambda \geq 0, H} \left\{ \phi_{\mathcal{T}}(\lambda) + \sigma^2 \mathrm{Tr}(H^T H) : \left[\begin{array}{c|c} \sum_k \lambda_k S_k & B^T - A^T H \\ \hline B - H^T A & I \end{array} \right] \succeq 0 \right\}$$

The efficiently computable optimal solution (λ_*, H_*) to this problem gives rise to the linear estimate

$$\widehat{x}_{H_*}(\omega) = H_*^T \omega$$

with risk not exceeding $\sqrt{\mathrm{Opt}}$. This estimate is near-optimal among all linear estimates:

$$\mathrm{Risk}[\widehat{x}_{H_*}|\mathcal{X}] \leq 2\sqrt{\ln(5K)} \cdot \inf_H \mathrm{Risk}[\widehat{x}_H|\mathcal{X}]$$
$$[\widehat{x}_H(\omega) = H^T \omega]$$

♠ **Surprising fact:** *The linear estimate $\widehat{x}_{H_*}$ is nearly optimal among all estimates, linear and nonlinear alike.*

♣ **Theorem.** *Let us associate with ellitope*

$$\mathcal{X} = \{x : \exists (t \in \mathcal{T}, y) : x = Ry, y^T S_k y \le t_k,\, k \le K\}$$

the convex compact set

$$\mathcal{Q} = \{Q \in \mathbf{S}^N : Q \succeq 0, \exists t \in \mathcal{T} : \mathrm{Tr}(S_k Q) \le t_k, k \le K\},$$

and the quantity

$$M_* = \max_{Q \in \mathcal{Q}} \sqrt{\mathrm{Tr}(BRQR^T B^T)}.$$

The linear estimate $\widehat{x}_{H_*}(\omega)$ *of* Bx, $x \in \mathcal{X}$, *via observation* $\omega = Ax + \sigma\xi, \xi \sim \mathcal{N}(0, I_m)$, *given by the optimal solution to the convex optimization problem*

$$\mathrm{Opt} = \min_{\lambda \ge 0, H} \left\{ \phi_{\mathcal{T}}(\lambda) + \sigma^2 \mathrm{Tr}(H^T H) : \left[\begin{array}{c|c} \sum_k \lambda_k S_k & B^T - A^T H \\ \hline B - H^T A & I \end{array}\right] \succeq 0 \right\}$$

satisfies the risk bound

$$\mathrm{Risk}[\widehat{x}_{H_*} | \mathcal{X}] \le \sqrt{\mathrm{Opt}} \le \sqrt{6 \ln\left(\frac{8 M_*^2 K}{\mathrm{Risk}^2_{\mathrm{opt}}[\mathcal{X}]}\right)} \mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}],$$

where

$$\mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}] = \inf_{\widehat{x}(\cdot)} \sup_{x \in \mathcal{X}} \sqrt{\mathbf{E}_{\xi \sim \mathcal{N}(0, I_m)} \left\{ \|Bx - \widehat{x}(Ax + \sigma\xi)\|_2^2 \right\}},$$

inf *being taken with respect to all, linear and nonlinear alike, estimates* $\widehat{x}(\cdot)$, *is the optimal minimax risk.*

♠ **Explanation, Easy part:** Consider the parametric convex optimization problem

$$\mathrm{Opt}(\rho) = \max_{Q\succeq 0, t\in\mathcal{T}} \left\{ \mathrm{Tr}\left(B\left[Q - QA^T(\sigma^2 I + AQA^T)^{-1}AQ\right]B^T\right) : \mathrm{Tr}(S_k Q) \leq \rho t_k, k \leq K \right\}$$

♡ **Objective:** Optimal expected $\|\cdot\|_2^2$-risk of recovery $B\eta$ from observation $A\eta+\sigma\xi$ with $\xi \sim \mathcal{N}(0, I)$ independent of *random Gaussian signal* $\eta \sim \mathcal{N}(0, Q)$.

♡ **Constraints** ensure that the probability for $\eta \sim \mathcal{N}(0, Q)$ *not* to belong to $\mathcal{X}$ goes to 0 exponentially fast (as $K\mathrm{e}^{-\frac{1}{3\rho}}$) as $\rho \to +0$

⇒ *Minimax optimal risk* $\mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}]$ *can be lower-bounded in terms of* $\mathrm{Opt}(\cdot)$

♠ **Explanation, Miracle part:** *By conic duality,* $\mathrm{Opt}(1)$ *turns out to be exactly the upper bound* Opt *on the squared risk of the near-optimal linear estimate, and by trivial reasons* $\mathrm{Opt}(\rho) \geq \rho\mathrm{Opt}(1)$

⇒ *Minimax optimal risk* $\mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}]$ *can be lower-bounded in terms of* Opt *and thus - in terms of the risk of the near-optimal linear estimate.*