Semidefinite Relaxation and Statistical Estimation

A. Nemirovski
Georgia Institute of Technology

joint research with

Anatoli Juditsky
Université Grenoble Alpes

Bridging Mathematical Optimization, Information Theory, and Data Science

Princeton Center for Theoretical Science

May 14-16, 2018

Problem of interest: *Given noisy observation*

$$\omega = Ax + \eta$$

- x – *unknown signal* known to belong to a given convex compact *signal set* $\mathcal{X} \subset \mathbb{R}^n$
- A – given $m \times n$ *sensing matrix*
- η – observation noise,

we want to recover linear image Bx *of the signal.*

- B – given $\nu \times n$ matrix.

Model of noise: η is random with the matrix of second moments

$$\mathsf{Var}[\eta] := \mathbf{E}\{\eta\eta^T\}$$

known to belong to a given convex compact subset $\mathbf{\Theta}$ of the cone $\mathbf{S}_+^m$ of positive semidefinite $m \times m$ matrices.

$$\boxed{\omega = Ax + \eta \;\&\; x \in \mathcal{X} \quad ?? \Rightarrow ?? \quad \widehat{x}(\omega) \approx Bx}$$

Performance of a candidate estimate $\widehat{x}(\omega) : \mathbb{R}^m \to \mathbb{R}^\nu$ is quantified by its $\|\cdot\|$-*risk*:

$$\mathrm{Risk}_{\|\cdot\|,\mathbf{\Theta}}[\widehat{x}|\mathcal{X}] = \sup_{x\in\mathcal{X},\eta:\mathrm{Var}[\eta]\in\mathbf{\Theta}} \mathbf{E}\left\{\|\widehat{x}(Ax+\eta) - Bx\|\right\}$$

- $\|\cdot\|$: a given norm on $\mathbb{R}^\nu$.

What we want: *efficiently computable, along with its risk bound, estimate with provably near-optimal performance*.

$$\boxed{\omega = Ax + \eta \;\&\; x \in \mathcal{X} \quad ?? \Rightarrow ?? \quad \widehat{x}(\omega) \approx Bx}$$

Motivation: In traditional Nonparametric Statistics, near-optimal estimates and their risks are yielded by *closed form analytical risk analysis.* Usually this analysis imposes severe restrictions on the data $A, B, \mathcal{X}$ and is problematic even in the

Simple case:

- $\eta \sim \mathcal{N}(0, \sigma^2 I)$ is white Gaussian noise,
- the recovery error is measured in ℓ_2-norm: $\|\cdot\| = \|\cdot\|_2$
- $\mathcal{X} = \{x \in \mathbb{R}^n : \sum_i a_i^2 x_i^2 \leq 1\}$ is an ellipsoid (with $a_i = i^\alpha$, $\mathcal{X}$ is comprised of discretizations of continuous time signals from a Sobolev ball).

$$\boxed{\omega = Ax + \eta \;\&\; x \in \mathcal{X} \quad ?? \Rightarrow ?? \quad \widehat{x}(\omega) \approx Bx}$$

Why analytical analysis is difficult: *optimal risk is determined by difficult to represent and analyze interplay between "geometries" of* $A, B, \mathcal{X}$.

— on one hand, poor conditioning of A *can make impossible* good recovery of some components of x even in a low noise

— on the other hand, the geometries of $\mathcal{X}$ and/or B can make "difficult to recover" components of x irrelevant – these components can be a priori small due to the geometry of the signal set $\mathcal{X}$, or can be suppressed by B due to the geometry of B.

Analytical study of Simple case is doable (and is basically complete) when A and B are diagonal matrices, and seems to be intractable for "general" A and/or B.

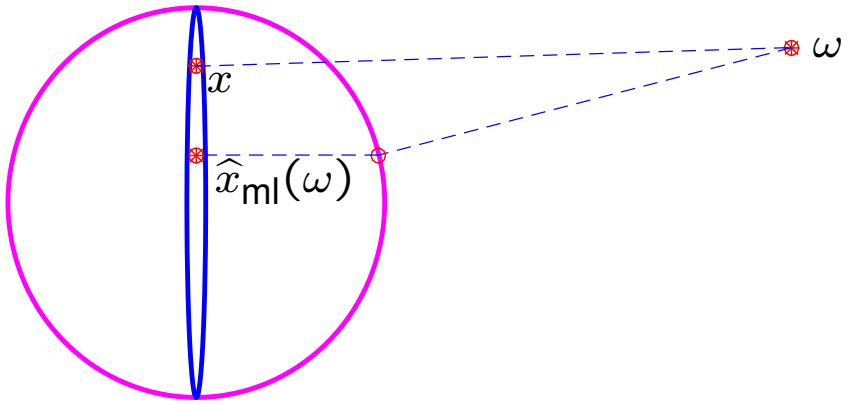
Surprisingly, difficulties disappear when passing from *analytical* to *computationally efficient numerical* design and risk analysis of estimates. Specifically, in Simple case *with no assumptions on* A, B (and even far beyond) one can build in a *computationally efficient* fashion *provably optimal*, up to logarithmic factors, estimates along with (upper bounds on) their risks.
Whether good or bad, these risks are nearly the best possible under the circumstances!

What is ahead: We intend to build two types of near minimax optimal, under appropriate assumptions on $\mathcal{X}$ and $\|\cdot\|$, estimates:

- *linear estimate* $\widehat{x}(\omega) = H^T\omega$
- *polyhedral estimate* $\widehat{x} = B \cdot \mathrm{argmin}_{x\in\mathcal{X}} \|H^T[\omega - Ax]\|_\infty$

$$\boxed{\omega = Ax + \eta \;\&\; x \in \mathcal{X} \quad ?? \Rightarrow ?? \quad \widehat{x}(\omega) \approx Bx}$$

Note: The statistical "magic wand" – Maximum Likelihood Estimate – can be *heavily nonoptimal* already in the Simple case.



blue: $\mathcal{X}$ magenta: $A\mathcal{X}$

- $\mathcal{X} = \{x \in \mathbb{R}^n : x_n^2 + \epsilon^{-2}\sum_{i=1}^{n-1} x_i^2 \le 1\}$
- $A = \mathrm{Diag}\{1/\epsilon, ..., 1/\epsilon, 1\}$, $\eta \sim \mathcal{N}(0, \sigma^2 I_n)$, $B = I_n$

$\Rightarrow$ MLE: $\widehat{x}_{\mathrm{ml}}(\omega) = A^{-1} \cdot \mathrm{argmin}_{\|u\|_2 \le 1} \|\omega - u\|_2$

When $\sigma \ll 1$, $\sigma^2 n \ge O(1)$, and $\epsilon \le O(\sigma)$, the risk of MLE is $O(1)$, while the minimax optimal risk is $O(\sigma)$.

$$\omega = Ax + \eta \;\&\; x \in \mathcal{X} \quad ?? \Rightarrow ?? \quad \widehat{x}_H(\omega) := H^T\omega \approx Bx$$

Near-Optimal Linear Estimation

What is known about linear estimates:

There is significant literature on linear estimates in minimax setting (primarily, with Gaussian signal-independent noise), including, but not reducing to,

Kuks&Olman '71,'72, Rao '72,'73, Pinsker '80, Efromovich&Pinsker '81,'82,'96, Pilz '81,'86, Donoho&Liu&McGibbon '90, Drygas '96, Christopeit&Helmes '96, Golubev&Levit&Tsybakov '96, Arnold&Stahlecker '00, Efromovich '08,...

This literature is mainly focused on design of "good" linear estimates.

However: Beyond

- Simple case with diagonal A, B (*Pinsker '80, Efromovich&Pinsker '96,...*) and its extension (A, B still diagonal, $\|\cdot\| = \|\cdot\|_2$, "quadratically convex" $\mathcal{X}$ rather than ellipsoid $\mathcal{X} = \{x \in \mathbb{R}^n : \sum_i a_i^2 x_i^2 \le 1\}$, *Donoho&Liu&McGibbon '90*)
- The case of $\mathrm{Rank}(B) = O(1)$ *with no restrictions on* A *and (convex compact)* $\mathcal{X}$ (*Donoho '94*)

we are *not* aware of preceding results on minimax near-optimality of properly built linear estimates.

Note: For "general" A, B and beyond Simple case, *already the design of near-optimal, within* $O(1)$ *factor, among all linear estimates seems to be computationally intractable*, except for the case of $\|\cdot\| = \|\cdot\|_\infty$.

$$\boxed{\omega = Ax + \eta \;\&\; x \in \mathcal{X} \quad ?? \Rightarrow ?? \quad \widehat{x}_H(\omega) := H^T\omega \approx Bx}$$

Fact: *Unless* $\mathrm{Rank}(B) = O(1)$*, linear estimates can be near-optimal only under some restrictions on the "geometry" of* $\mathcal{X}, \|\cdot\|$.

Example: When recovering signal $x \in \mathcal{X} := \{x \in \mathbb{R}^n : \|x\|_1 \leq 1\}$ from direct observations

$$\omega = x + \eta,\ \eta \sim \mathcal{N}(0, \sigma^2 I_n)$$

linear estimates are heavily nonoptimal: when $1 \geq \sigma \geq n^{-1/2}$,

- the best $\|\cdot\|_2$-risk achievable with linear estimates is $O(1)$
- the minimax optimal risk is $O(1)\sqrt{\sigma}$

Our standing assumptions when speaking about linear estimates are:

- The set Θ of allowed matrices of noise's second moments is a convex compact subset of the *interior* of $\mathbf{S}^m_+$
- $\mathcal{X}$ and the unit ball $\mathcal{B}_* = \{u : \|u\|_* \leq 1\}$ of the norm *conjugate* to $\|\cdot\|$:
$$\|u\|_* = \max\{u^T v : \|v\| \leq 1\},$$
are *ellitopes* or *spectratopes*.

Ellitopes

Basic ellitope in $\mathbb{R}^N$ is a *bounded* set $\mathcal{Z}$ given by representation

$$\mathcal{Z} = \{z \in \mathbb{R}^N : \exists t \in \mathcal{T} : z^T S_k z \le t_k,\ k \le K\}$$

where

- $S_k \succeq 0,\ k \le K$
- $\mathcal{T} \subset \mathbb{R}^K_+$ is convex compact set which contains a positive vector and is *monotone*: $0 \le t' \le t \in \mathcal{T}$ implies that $t' \in \mathcal{T}$.

Examples:

A. *Bounded* intersection of K ellipsoids/elliptic cylinders $\{x : x^T S_k x \le 1\}$ (set $\mathcal{T} = [0,1]^K$)

B. $\|\cdot\|_p$-norm ball, $2 \le p \le \infty$:

$$\{z \in \mathbb{R}^N, \|z\|_p \le 1\} = \{z \in \mathbb{R}^N : \exists t \in \mathcal{T} : z^T S_k z \equiv z_k^2 \le t_k, k \le N\},$$
$$\mathcal{T} = \{t \in \mathbb{R}^N_+ : \|t\|_{p/2} \le 1\}$$

Ellitope $\mathcal{X}$ is a set represented as linear image of a basic ellitope:

$$\mathcal{X} = \{x : \exists z \in \mathcal{Z} : x = Pz\} \text{ with } \mathcal{Z} = \{z \in \mathbb{R}^N : \exists t \in \mathcal{T} : z^T S_k z \le t_k,\ k \le K\}$$

Spectratopes

Basic spectratope in $\mathbb{R}^N$ is a *bounded* set $\mathcal{Z}$ given by representation

$$\mathcal{Z} = \{z \in \mathbb{R}^N : \exists t \in \mathcal{T} : S_k^2[z] \preceq t_k I_{\mu_k},\ k \le K\}$$

where

- $S_k[z] = \sum_{j=1}^N z_j S^{kj}$ is a $\mu_k \times \mu_k$ *symmetric* matrix linearly depending on z
- $\mathcal{T} \subset \mathbb{R}^K_+$ is as in the definition of ellitope.

Example: Matrix box $\{z \in \mathbb{R}^{p\times q} : \|z\|_{2,2} \le 1\}$ ($\|\cdot\|_{2,2}$ – spectral norm):

$$\{z \in \mathbb{R}^{p\times q} : \|z\|_{2,2} \le 1\} = \{z \in \mathbb{R}^{p\times q} : \exists t \in [0,1] : \left[\begin{array}{c|c} & z \\ \hline z^T & \end{array}\right]^2 \preceq tI_{p+q}\}.$$

Spectratope $\mathcal{X}$ is a set represented as linear image of a basic spectratope:

$$\mathcal{X} = \{x : \exists z \in \mathcal{Z} : x = Pz\},\ \mathcal{Z} = \{z \in \mathbb{R}^N : \exists t \in \mathcal{T} : S_k^2[z] \preceq t_k I_{\mu_k},\ k \le K\}$$

Fact: *Every ellitope is a spectratope.*

Fact: *Ellitopes/Spectratopes admit fully algorithmic calculus:* nearly all operations preserving "built-in" properties of these sets – convexity, compactness and symmetry w.r.t. the origin, like taking

- finite intersections,
- direct products,
- arithmetic sums,
- linear images,
- inverse images under linear embeddings,

as applied to ellitopes/spectratopes, result in sets of the same type, with ellitopic /spectratopic representation readily given by representations of the operands.

- *What is missing, is taking convex hull of finite union*.

Fact: *Norms $\|\cdot\|$ with ellitopes/spectratopes as the unit balls of their duals $\|\cdot\|_*$ include $\|\cdot\|_p$, $1 \le p \le 2$, and nuclear norm. The family of these norms admits fully algorithmic calculus and is closed w.r.t. basic operations like*

- summation with positive coefficients,
- direct summation $\{\|\cdot\|_{(i)}, i \le I\} \mapsto \|[x^1; ...; x^I]\| = \sum_i \|x^i\|_{(i)}$,
- superpositions with linear embeddings,
- passing to factor-norms.

- *What is missing, is taking maximum*.

$$\boxed{\omega = Ax + \eta \;\&\; x \in \mathcal{X} \quad ?? \Rightarrow ?? \quad \widehat{x}_H(\omega) := H^T\omega \approx Bx}$$

Building presumably good linear estimate

Immediate observation: The $\|\cdot\|$-risk of a linear estimate $\widehat{x}_H(\omega) := H^T\omega$ can be tightly upper-bounded as

$$\mathrm{Risk}_{\|\cdot\|}[\widehat{x}_H|\mathcal{X}] \le \underbrace{\max_{x\in\mathcal{X}} \|[B - H^TA]x\|}_{\text{bias } \Phi^*_{\mathcal{X}}(H)} + \underbrace{\max_{\eta:\mathrm{Var}[\eta]\in\mathbf{\Theta}} \mathbf{E}_\eta\left\{\|H^T\eta\|\right\}}_{\text{stochastic term } \Psi^*_{\mathbf{\Theta}}(H)}$$

$\Rightarrow$ *A nearly ideal linear estimate is yielded by*

$$H_* \in \mathrm{Argmin}_H\left[\Phi^*_{\mathcal{X}}(H) + \Psi^*_{\mathbf{\Theta}}(H)\right]$$

But: $\Phi^*_{\mathcal{X}}(\cdot)$ and $\Psi^*_{\mathbf{\Theta}}(\cdot)$, while convex, can be difficult to compute...

Remedy: To get "presumably good" linear estimate, use the optimal solution to the efficiently solvable problem

$$\mathrm{Opt} = \min_H\left[\Phi(H) + \Psi(H)\right]$$

where $\Phi(H)$, $\Psi(H)$ are *efficiently computable convex upper bounds* on $\Phi^*_{\mathcal{X}}(H)$, $\Psi^*_{\mathbf{\Theta}}(H)$.

Tight upper bounding of bias $\Phi^*_{\mathcal{X}}(H) := \max\limits_{x\in\mathcal{X}} \|[B - H^T A]x\|$

Immediate observation: *$\Phi^*_{\mathcal{X}}(H)$ is the maximum of affinely parameterized by H quadratic form over the set* $\underbrace{\{u : \|u\|_* \leq 1\}}_{\mathcal{B}_*} \times \mathcal{X}$*:*

$$\Phi^*_{\mathcal{X}}(H) = \max_{[u;x]\in\mathcal{B}_*\times\mathcal{X}} u^T[B - H^T A]x = \tfrac{1}{2} \max_{[u;x]\in\mathcal{B}_*\times\mathcal{X}} [u;x]^T \left[\begin{array}{c|c} & B - H^T A \\ \hline B^T - A^T H & \end{array}\right] [u;x].$$

Fact: When $\mathcal{B}_*$ and $\mathcal{X}$ are spectratopes, so is $\mathcal{Y} = \mathcal{B}_* \times \mathcal{X}$, and *a quadratic form $y^T C y$ on a spectratope can be tightly upper-bounded, in a computationally efficient fashion, via semidefinite relaxation*.

Disclaimer: To save notation, from now on $\mathcal{X}$ and $\mathcal{B}_*$ are assumed to be *basic* ellitopes/spectratopes. Extensions to general ellitopes/spectratopes are straightforward.

In the ellitopic case, SD relaxation is given by

Theorem [Ju&N'16] *Let*

$$\mathcal{Z} = \{z : \exists t \in \mathcal{T} : z^T S_k z \le t_k, k \le K\} \subset \mathbb{R}^N$$

be an ellitope, and let $C \in \mathbf{S}^N$. *Then*

- [Easy part] *The optimal value in the convex optimization problem*

$$\begin{array}{c}\mathrm{Opt}(C) = \min_\lambda \{\phi_{\mathcal{T}}(\lambda) : \lambda \ge 0, C \preceq \sum_k \lambda_k S_k\} \\ \phi_{\mathcal{T}}(\lambda) = \max_{t\in\mathcal{T}} t^T\lambda : \text{ support function of } \mathcal{T}\end{array}$$

is an efficiently computable convex in C *upper bound on* $\mathcal{M}(C) = \max\limits_{z\in\mathcal{Z}} z^T C z$.

- [Not so easy part] *The above bound is reasonably tight:*

$$\mathcal{M}(C) \le \mathrm{Opt}(C) \le 3\ln(\sqrt{3}K)\mathcal{M}(C).$$

Proof of Easy part: When $\lambda \ge 0$ and $C \preceq \sum_k \lambda_k S_k$, we have

$$\begin{array}{rcl}\mathcal{M}(C) & = & \max\limits_{z\in\mathcal{Z}} z^T C z \\ & \le & \max\limits_{z\in\mathcal{Z}} z^T[\sum_k \lambda_k S_K]z = \max\limits_{t\in\mathcal{T},z} \{\sum_k \lambda_k z^T S_k z : z^T S_k z \le t_k, k \le K\} \\ & \le & \max\limits_{t\in\mathcal{T}} \{\sum_k \lambda_k t_k\} = \phi_{\mathcal{T}}(\lambda).\end{array}$$

Not so easy part: sketch of the proof:

A. Invoking Conic Duality, we get

$$\begin{array}{rcl}\mathrm{Opt}(C) & = & \max\limits_{Z,t}\{\mathrm{Tr}(CZ) : t\in\mathcal{T}, Z\succeq 0, \mathrm{Tr}(S_kZ)\le t_k, k\le K\}\\ & = & \mathrm{Tr}(CZ_*) \quad [Z_*\succeq 0, \exists t^*\in\mathcal{T} : \mathrm{Tr}(S_kZ_*)\le t^*_k, k\le K]\end{array}$$

B. $Z_*\succeq 0 \Rightarrow Z_* = \mathbf{E}\{\zeta\zeta^T\}$ for properly selected *random* ζ

Note: $\mathbf{E}\{\zeta^TS_k\zeta\}\le t^*_k, k\le K$ and $\mathbf{E}\{\zeta^TC\zeta\} = \mathrm{Opt}(C)$

C. On a close inspection, ζ *can be built in such a way that*

(a) $\zeta^TC\zeta\equiv\mathrm{Opt}(C)$

(b) ζ *is "light-tail," so that*

$$\mathrm{Prob}\left\{\zeta^TS_k\zeta > \theta\mathbf{E}\{\zeta^TS_k\zeta\}\right\}\le\sqrt{3}\mathrm{e}^{-\theta/3}\ \forall(k\le K,\theta\ge 0)$$

$\Rightarrow$ *some realization* $\bar\zeta$ *of* ζ *satisfies*

$$\bar\zeta^TS_k\bar\zeta\le 3\ln(\sqrt{3}K)t^*_k\,\forall k\ \&\ \bar\zeta^TC\bar\zeta = \mathrm{Opt}(C)$$

$\Rightarrow$ $z = \bar\zeta/\sqrt{3\ln(\sqrt{3}K)}$ *is a feasible solution to the problem* $\mathcal{M}(C) = \max\limits_{z\in\mathcal{Z}} z^TCz$ *with the value of the objective* $\ge \mathrm{Opt}(C)/(3\ln(\sqrt{3}K))$. □

In the spectratopic case, SD relaxation is given by

Theorem [Ju&N'17] *Let*

$$\mathcal{Z} = \{z : \exists t \in \mathcal{T} : S_k^2[z] \preceq t_k I_{\mu_k}, k \le K\} \subset \mathbb{R}^N$$
$$\left[S_k[z] = \textstyle\sum_j z_j S^{kj} : \mathbb{R}^N \to \mathbf{S}^{\mu_k}\right]$$

be a spectratope, and let $C \in \mathbf{S}^N$. *Then*

- [Easy part] *The optimal value in the convex optimization problem*

$$\mathrm{Opt}(C) = \min_{\Lambda_1,\dots,\Lambda_K} \left\{ \phi_{\mathcal{T}}([\mathrm{Tr}(\Lambda_1);\dots;\mathrm{Tr}(\Lambda_K)]) : \begin{array}{c} 0 \preceq \Lambda_k \in \mathbf{S}^{\mu_k}, k \le K \\ C \preceq \sum_k \mathcal{S}_k(\Lambda_k) \end{array} \right\} \qquad (*)$$
$$\left[\phi_{\mathcal{T}}(\lambda) = \max_{t\in\mathcal{T}} t^T\lambda, \mathcal{S}_k(\Lambda_k) : \mathbf{S}^{\mu_k} \mapsto \mathbf{S}^N : [\mathcal{S}_k(\Lambda_k)]_{ij} = \mathrm{Tr}\left(S^{ki}\Lambda_k S^{kj}\right)\right]$$

is an efficiently computable convex upper bound on $\mathcal{M}(C) = \max_{z\in\mathcal{Z}} z^T C z$.

- [Not so easy part] *The above bound is reasonably tight:*

$$\mathcal{M}(C) \le \mathrm{Opt}(C) \le 2\ln(3\textstyle\sum_k \mu_k)\mathcal{M}(C).$$

Proof of Easy part: When Λ_k, $k \le K$ are feasible for $(*)$ and $z \in \mathcal{Z}$, we have

$$\begin{array}{rcl} z^T C z & \le & z^T[\sum_k \mathcal{S}_k(\Lambda_k)]z = \sum_k \sum_{i,j} \mathrm{Tr}(S^{ki}\Lambda_k S^{kj}) z_i z_j \\ & = & \sum_k \mathrm{Tr}\left(S_k[z]\Lambda_k S_k[z]\right) = \sum_k \mathrm{Tr}\left(\Lambda_k S_k^2[z]\right) \le \max_{t\in\mathcal{T}} \sum_k \mathrm{Tr}(\Lambda_k t_k I_{\mu_k}) \\ & \le & \max_{t\in\mathcal{T}} \sum_k t_k \mathrm{Tr}(\Lambda_k) = \phi_{\mathcal{T}}([\mathrm{Tr}(\Lambda_1);\dots;\mathrm{Tr}(\Lambda_K)]) \end{array}$$

Tight upper bounding of stochastic term $\Psi^*_{\Theta}(H) := \sup_{\eta:\mathsf{Var}[\eta]\in\Theta} \mathbf{E}_\eta\{\|H^T\eta\|\}$

Theorem [Ju&N'17] *Let the unit ball of $\|\cdot\|_*$ be a spectratope:*

$$\{u\in\mathbb{R}^\nu : \|u\|_*\le 1\} = \{u : \exists p\in\mathcal{P} : R_\ell^2[u]\preceq p_\ell I_{\nu_\ell}, \ell\le L\}$$
$$\left[R_\ell[u] = \textstyle\sum_j u_j R^{\ell j} : \mathbb{R}^\nu \to \mathbf{S}^{\nu_\ell}\right]$$

and let Θ be a convex compact subset of $\mathsf{int}\,\mathbf{S}^\nu_+$. Then

- [Easy part] *The optimal value in the explicit convex optimization problem*

$$\Psi_\Theta(U) = \min_{\Upsilon,G}\left\{\phi_{\mathcal{P}}([\mathsf{Tr}(\Upsilon_1);...;\mathsf{Tr}(\Upsilon_L)]) + \max_{\Sigma\in\Theta}\mathsf{Tr}(\Sigma G) : \begin{array}{c} 0\preceq\Upsilon_\ell\in\mathbf{S}^{\nu_\ell},\ell\le L \\ \left[\begin{array}{c|c} G & \frac{1}{2}U \\ \hline \frac{1}{2}U^T & \sum_\ell \mathcal{R}_\ell(\Upsilon_\ell)\end{array}\right]\succeq 0\end{array}\right\}$$
$$\left[\phi_{\mathcal{P}}(\lambda) = \max_{p\in\mathcal{P}} p^T\lambda,\ \mathcal{R}_\ell(\Upsilon_\ell):\mathbf{S}^{\nu_\ell}\to\mathbf{S}^\nu : [\mathcal{R}_\ell(\Upsilon_\ell)]_{ij} = \mathsf{Tr}\left(R^{\ell i}\Upsilon_\ell R^{\ell j}\right)\right]$$

is an efficiently computable convex upper bound on $\Psi^*_\Theta(U) = \sup_{\eta:\mathsf{Var}[\eta]\in\Theta}\mathbf{E}_\eta\{\|U^T\eta\|\}$.

- [Not so easy part] *The upper bound $\Psi_\Theta(U)$ on $\mathbf{E}_\eta\{\|U^T\eta\|\}$ is reasonably tight already for zero mean Gaussian η:*

$$\Psi^*_\Theta(U)\le\Psi_\Theta(U)\le 31\sqrt{\ln(44\textstyle\sum_\ell\nu_\ell)}\max_{\Sigma\in\Theta}\mathbf{E}_{\eta\sim\mathcal{N}(0,\Sigma)}\{\|U^T\eta\|\}.$$

Illustration: *Let* $\|\cdot\| = \|\cdot\|_p$ *with* $1 \le p \le 2$ *and* $\Theta = \{\theta\}$. *Then the upper bound* $\Psi_{\{\Sigma\}}(H)$ *on* $\mathbf{E}\{\|H^T\eta\|_p\}$, $\mathsf{Var}[\eta] = \Sigma$, *becomes*

$$\Psi_{\{\Sigma\}}(H) = \left\|\left[\|\Sigma^{1/2}\mathsf{Col}_1[H]\|_2; ...; \|\Sigma^{1/2}\mathsf{Col}_\nu[H]\|_2\right]\right\|_p$$

Proof of Easy part: When Υ, G is feasible for optimization problem specifying $\Psi_\Theta(U)$, we have

$$0 \preceq \Upsilon_\ell,\ \ell \le L \ \&\ \left[\begin{array}{c|c} G & \frac{1}{2}U \\ \hline \frac{1}{2}U^T & \sum_\ell \mathcal{R}_\ell(\Upsilon_\ell) \end{array}\right] \succeq 0,$$

whence

$$\begin{array}{rcl} \|U^T\eta\| & = & \max\limits_{u:\|u\|_*\le 1} \eta^T U u \le \max_{\|u\|_*\le 1}\left[\eta^T G\eta + u^T[\sum_\ell \mathcal{R}_\ell(\Upsilon_\ell)]u\right] \\ & \le & \eta^T G\eta + \phi_{\mathcal{P}}\left([\mathsf{Tr}(\Upsilon_1); ...; \mathsf{Tr}(\Upsilon_L)]\right) \\ \Rightarrow & & \mathbf{E}_\eta\{\|U^T\eta\|\} \le \mathsf{Tr}(G\,\mathsf{Var}[\eta]) + \phi_{\mathcal{P}}\left([\mathsf{Tr}(\Upsilon_1); ...; \mathsf{Tr}(\Upsilon_L)]\right) \qquad \square \end{array}$$

$$\boxed{\omega = Ax + \eta \;\&\; x \in \mathcal{X} \quad ?? \Rightarrow ?? \quad \widehat{x}_H(\omega) := H^T\omega \approx Bx}$$

Assembling the blocks

Bottom Line *Let the signal set $\mathcal{X}$ and the unit ball $\mathcal{B}_*$ of norm $\|\cdot\|_*$ be spectratopes:*

$$\begin{array}{rcl}\mathcal{X} & = & \{x \in \mathbb{R}^n : \exists t \in \mathcal{T}, z : x = Pz \;\&\; S_k^2[z] \preceq t_k I_{\mu_k}, k \le K\} \quad [S_k[z] = \sum_j z_j S^{kj}] \\ \mathcal{B}_* & = & \{u \in \mathbb{R}^\nu : \exists p \in \mathcal{P}, w : u = Qw \;\&\; R_\ell^2[w] \preceq p_\ell I_{\nu_\ell}, \ell \le L\} \quad [R_\ell[w] = \sum_j w_j R^{\ell j}]\end{array}$$

Then the linear estimate $\widehat{x}_{H_*} = H_*^T\omega$ yielded by an optimal solution to the efficiently solvable convex optimization problem

$$\mathsf{Opt}(P) = \min_H \{\Phi(H) + \Psi(H)\}, \qquad (P)$$

($\Phi(H), \Psi(H)$ are the already built efficiently computable convex upper bounds on the bias and the stochastic term in the risk of a linear estimate $\omega \mapsto H^T\omega$) satisfies the risk bound

$$\mathsf{Risk}_{\|\cdot\|}[\widehat{x}_{H_*}|\mathcal{X}] := \sup_{\eta:\mathsf{Var}[\eta]\in\mathbf{\Theta}} \max_{x\in\mathcal{X}} \mathbf{E}_\eta\left\{\|Bx - H_*^T[Ax+\eta]\|\right\} \le \mathsf{Opt}(P)$$

and is the best, within factor $O(1)[\ln(\sum_k \mu_k) + \ln(\sum_\ell \nu_\ell)]$, in terms of $\|\cdot\|$-risk among all linear estimates.

$$\omega = Ax + \eta \;\&\; x \in \mathcal{X} \quad ?? \Rightarrow ?? \quad \widehat{x}_H(\omega) := H^T\omega \approx Bx$$
$$\mathcal{X} = \{x \in \mathbb{R}^n : \exists t \in \mathcal{T} : S_k^2[x] \preceq t_k I_{\mu_k}, k \le K\}$$
$$\mathcal{B}_* := \{u : \|u\|_* \le 1\} = \{u \in \mathbb{R}^\nu : \exists p \in \mathcal{P} : R_\ell^2[u] \preceq p_\ell I_{\nu_\ell}, \ell \le L\}$$

Main Result [Ju&N'17] *Under the premise of Bottom Line, the linear estimate $\widehat{x}_{H_*}$ is near-optimal among all estimates:*

$$\mathrm{Risk}_{\|\cdot\|}[\widehat{x}_{H_*}|\mathcal{X}] \le \mathrm{Opt}(P) \le O(1)\sqrt{\ln(2\mathfrak{n})\ln\left(\frac{\mathfrak{s}\sqrt{\mathfrak{m}}\|B\|_{2,2}D[\mathcal{X}]}{\mathrm{RiskOpt}_{\|\cdot\|}[\mathcal{X}]}\right)}\,\mathrm{RiskOpt}_{\|\cdot\|}[\mathcal{X}]$$

- $\|\cdot\|_{2,2}$: spectral norm
- $D[\mathcal{X}]$: $\|\cdot\|_2$-diameter of $\mathcal{X}$
- $\mathfrak{s} = \max\limits_{\|u\|_2\le 1}\|u\|$
- $\mathfrak{m} = \sum_k \mu_k$, $\mathfrak{n} = \sum_\ell \nu_\ell$

$\mathrm{RiskOpt}_{\|\cdot\|}[\mathcal{X}] = \inf\limits_{\widehat{x}(\cdot)} \sup\limits_{\Sigma\in\mathbf{\Theta},x\in\mathcal{X}} \mathbf{E}_{\eta\sim\mathcal{N}(0,\Sigma)}\{\|Bx - \widehat{x}(Ax+\eta)\|\}$: minimax optimal Gaussian risk

Surprise: *Logarithmic "tightness factor" in (!) is not directly affected by the geometries of $\mathcal{X}$, $\|\cdot\|$, A and B – the entities primarily responsible for the minimax optimal risk.*

Proof Strategy

- By simple saddle point argument, the upper bound Opt on the risk of the linear estimate $\widehat{x}_{H_*}$ remains intact when the set $\mathbf{\Theta}$ of allowed noise moment matrices is replaced with properly selected singleton $\{\Sigma\} \subset \mathbf{\Theta}$. We restrict the observation noise to be $\mathcal{N}(0, \Sigma)$. Besides this, we lose nothing when assuming $\mathcal{X}$ to be a *basic* spectratope.
- The idea of the proof goes back to M. Pinsker '80 who considered the Simple case with direct observations $A = B = I$ and $\|\cdot\| = \|\cdot\|_2$. Given $W \succeq 0$, consider the optimal *Bayesian* risk

$$\mathrm{RiskB}[W] = \inf_{\widehat{x}(\cdot)} \mathbf{E}_{[\eta;\xi]\sim\mathcal{N}(0,\Sigma)\times\mathcal{N}(0,W)} \left\{\|B\xi - \widehat{x}(A\xi+\eta)\|\right\},$$

- Similarly to the Gauss-Markov Theorem, it is easily seen that the optimal Bayesian risk is "nearly achieved" at a *linear* estimate:

$$\Gamma^*(W) := \min_H \left\{\mathbf{E}_{\xi\sim\mathcal{N}(0,W)}\left\{\|[B - H^T A]\xi\|\right\} + \mathbf{E}_{\eta\sim\mathcal{N}(0,\Sigma)}\left\{\|H^T\eta\|\right\}\right\} \le O(1)\mathrm{RiskB}[W].$$

This combines with Theorem on tightness of the upper bound $\Psi_{\{S\}}(F)$ on $\mathbf{E}_{\zeta\sim\mathcal{N}(0,S)}\left\{\|F^T\zeta\|\right\}$ to imply

$$\forall W \succeq 0 : \Gamma(W) := \min_H \left\{\Psi_{\{W\}}(B - H^T A) + \Psi_{\{\Sigma\}}(H)\right\} \le O(1)\sqrt{\ln(\mathfrak{n})}\mathrm{RiskB}[W]. \qquad (1)$$

- Heavily exploiting conic duality, we build a "path" $W_\rho \succeq 0$, $\rho \in [0, 1]$ such that
  - $\sqrt{\rho}\,\mathrm{Opt} \le \Gamma(W_\rho)$, $0 < \rho \le 1$ (2)
  - $\mathrm{Prob}_{\xi\sim\mathcal{N}(0,W_\rho)}\{\xi \notin \mathcal{X}\} \le O(1)\mathfrak{m}\exp\{-O(1)/\rho\}$

  $\Rightarrow$ $\mathrm{RiskB}[W_\rho] \le O(1)\left[\mathfrak{s}\sqrt{\mathfrak{m}}\|B\|_{2,2}D[\mathcal{X}]\exp\{-O(1)/\rho\} + \mathrm{RiskOpt}_{\|\cdot\|}[\mathcal{X}]\right]$ (3)
- (1) – (3) with properly selected ρ results in the upper bound on Opt via RiskOpt from Main Result.

Variation: Uncertain-but-Bounded noise

Situation: Given observation $\omega = Ax + \eta$ of *unknown* signal x *known to belong to a given signal set* $\mathcal{X}$, we want to recover Bx. All we know about the noise is $\eta \in \mathcal{H}$, with a known and bounded set $\mathcal{H}$.

- We define the risk of an estimate $\omega \mapsto \widehat{x}(\omega)$ as

$$\mathrm{Risk}_{\|\cdot\|,\mathcal{H}}[\widehat{x}|\mathcal{X}] = \sup_{x\in\mathcal{X},\eta\in\mathcal{H}} \|Bx - \widehat{x}(Ax+\eta)\|$$

Assumption: *$\mathcal{X}, \mathcal{H}$ are spectratopes, and the unit ball of $\|\cdot\|_*$ is a basic spectratope*

Observation: *By redefining signal as* $[x;\eta]$ *and replacing* $A \leftarrow [A, I]$, $B \leftarrow [B, 0]$, $\mathcal{X} \leftarrow \mathcal{X} \times \mathcal{H}$, *the problem reduces to the problem*

Given spectratope

$$\mathcal{X} = \{x \in \mathbb{R}^n : \exists t \in \mathcal{T} : S_k^2[x] \preceq t_k I_{\mu_k}, k \leq K\}$$

and norm $\|\cdot\|$ *such that the unit ball of* $\|\cdot\|_*$ *is a basic spectratope:*

$$\{u : \|u\|_* \leq 1\} = \{u : \exists p \in \mathcal{P} : R_\ell^2[u] \preceq p_\ell I_{\nu_\ell}, \ell \leq L\},$$

recover Bx *with unknown* $x \in \mathcal{X}$ *from observation* $\omega = Ax$.

Fact: *In the above situation, an efficiently computable linear estimate* $\widehat{x}(\omega) = H^T\omega$ *is minimax optimal within a logarithmic factor.*

$$\omega = Ax \;\&\; x \in \mathcal{X} \quad ?? \Rightarrow ?? \quad \widehat{x}_H(\omega) := H^T\omega \approx Bx$$
$$\mathcal{X} = \{x \in \mathbb{R}^n : \exists t \in \mathcal{T} : S_k^2[x] \preceq t_k I_{\mu_k}, k \le K\} \;\; [S_k[x] = \textstyle\sum_j z_j S^{kj}]$$
$$\mathcal{B}_* := \{u : \|u\|_* \le 1\} = \{u \in \mathbb{R}^\nu : \exists p \in \mathcal{P} : R_\ell^2[w] \preceq p_\ell I_{\nu_\ell}, \ell \le L\} \;\; [R_\ell[u] = \textstyle\sum_j u_j R^{\ell j}]$$

Observation: Risk of a linear estimate $\widehat{x}_H(\omega) = H^T\omega$ is the maximum of a quadratic form, affinely parameterized by H, over the spectratope $\mathcal{B}_* \times \mathcal{X}$:

$$\mathrm{Risk}_{\|\cdot\|}[\widehat{x}_H|\mathcal{X}] := \max_{x\in\mathcal{X}} \|[B - H^TA]x\| = \max_{[u;x]\in\mathcal{B}_*\times\mathcal{X}} [u;x]^T \left[\begin{array}{c|c} & \frac{1}{2}[B - H^TA] \\ \hline \frac{1}{2}[B^T - A^TH] & \end{array}\right] [u;x]$$

$\Rightarrow$ [SD relaxation] *"Presumably good" linear estimate $\widehat{x}_{H_*}(\omega) = H_*^T\omega$ given by the H-component of optimal solution to the problem*

$$\mathrm{Opt} = \min_{\Lambda,\Upsilon} \left\{ \phi_{\mathcal{T}}([\mathrm{Tr}(\Lambda_1); ...; \mathrm{Tr}(\Lambda_K)]) + \phi_{\mathcal{P}}([\mathrm{Tr}(\Upsilon_1); ...; \mathrm{Tr}(\Upsilon_L)]) : \right.$$
$$\left. 0 \preceq \Lambda_k \in \mathbf{S}^{\mu_k}, k \le K, 0 \preceq \Upsilon_\ell \in \mathbf{S}^{\nu_\ell} \;\; \left[\begin{array}{c|c} \sum_\ell \mathcal{R}_\ell(\Upsilon_\ell) & \frac{1}{2}[B - H^TA] \\ \hline \frac{1}{2}[B^T - A^TH] & \sum_k \mathcal{S}_k(\Lambda_k) \end{array}\right] \succeq 0 \right\}$$
$$\left[\begin{array}{c} \phi_{\mathcal{T}}(\lambda) = \max_{t\in\mathcal{T}} t^T\lambda, \phi_{\mathcal{P}}(\lambda) = \max_{p\in\mathcal{P}} p^T\lambda \\ [\mathcal{S}_k(\Lambda_k)]_{ij} = \mathrm{Tr}\left(S^{ki}\Lambda_k S^{kj}\right), [\mathcal{R}_\ell(\Upsilon_\ell)]_{ij} = \mathrm{Tr}\left(R^{\ell i}\Upsilon_\ell R^{\ell j}\right) \end{array} \right]$$

satisfies the risk bound $\mathrm{Risk}_{\|\cdot\|}[\widehat{x}_{H_*}|\mathcal{X}] \le \mathrm{Opt}$.

Theorem [Ju&N'17] *The linear estimate $\widehat{x}_{H_*}(\cdot)$ is nearly minimax optimal:*

$$\mathrm{Risk}_{\|\cdot\|}[\widehat{x}_{H_*}|\mathcal{X}] \le \mathrm{Opt} \le O(1)\left[\ln(\textstyle\sum_k \mu_k) + \ln(\sum_\ell \nu_\ell)\right] \inf_{\widehat{x}(\cdot)} \sup_{x\in\mathcal{X}} \|Bx - \widehat{x}(Ax)\|.$$

Note: In contrast to the case of random noise, the logarithmic "tightness factor" now is *not* affected by the minimax risk!

$$\boxed{\omega = Ax + \eta \;\&\; x \in \mathcal{X} \quad ?? \Rightarrow ?? \quad \widehat{x}(\omega) \approx Bx}$$

Near-optimal Polyhedral estimates

Disclaimer: For the sake of simplicity, we from now on restrict ourselves with *Sub-Gaussian*, with parameters $(0, \sigma^2 I)$ noise:

$$\eta \sim \mathcal{SG}_\sigma \Leftrightarrow \mathbf{E}_\eta\left\{\exp\{h^T\eta\}\right\} \leq \frac{\sigma^2}{2} h^T h \;\forall h$$

However: Constructions and results to follow can be extended to the cases of

- *Poisson observation scheme*, where the entries in ω are independent of each other Poisson random variables with parameters affinely parameterized by $x \in \mathcal{X}$
- *Discrete observation scheme* where ω is a K-element sample $\{\omega_t, t \leq K\}$ drawn, independently across t, from a finite support probability distribution affinely parameterized by $x \in \mathcal{X}$.

$$\boxed{\omega = Ax + \eta \;\&\; x \in \mathcal{X} \quad ?? \Rightarrow ?? \quad \widehat{x}(\omega) \approx Bx}$$

When speaking about polyhedral estimates, it is convenient to fix a "confidence tolerance" $\epsilon \in (0, 1)$ and to pass from $\|\cdot\|$-risk of an estimate $\widehat{x}(\cdot)$ to its $(\|\cdot\|, \epsilon)$-risk

$$\mathsf{Risk}_{\|\cdot\|,\epsilon}[\widehat{x}|\mathcal{X}] = \inf\{\rho : \mathsf{Prob}\{\eta : \|Bx - \widehat{x}(Ax + \eta)\| > \rho\} \le \epsilon \forall (x \in \mathcal{X}, \eta \sim \mathcal{SG}_\sigma)\}$$

– the worst-case, w.r.t. $x \in \mathcal{X}, \eta \sim \mathcal{SG}_\sigma$, $\|\cdot\|$-size of the $(1-\epsilon)$-confidence "interval" of the estimate.

$$\boxed{\mathbb{R}^m \ni \omega = Ax + \eta \;\&\; \eta \sim \mathcal{SG}_\sigma \;\&\; x \in \mathcal{X} \quad ?? \Rightarrow ?? \quad \widehat{x}(\omega) \approx Bx \in \mathbb{R}^\nu}$$

Immediate observation: *It is easy to estimate linear forms h^TAx of x: the "plug-in" estimate $\widehat{h}(\omega) = h^T\omega$ is unbiased and satisfies*

$$\forall x : \mathrm{Prob}_{\eta\sim\mathcal{SG}_\sigma}\left\{|\widehat{h}(Ax+\eta) - h^TAx| > \sigma\|h\|_2\sqrt{2\ln(2/\epsilon)}\right\} \le \epsilon$$

The simplest way to utilize Immediate observation when recovering Bx is offered by *generic polyhedral estimate*:

- Build somehow $m \times N$ *contrast matrix* H with $\|\cdot\|_2$-unit columns
- Given observation ω, find an optimal solution $x^H(\omega)$ to the convex problem

$$\min_u \left\{\|H^T(Au-\omega)\|_\infty : u \in \mathcal{X}\right\}$$

and take $\widehat{x}^H(\omega) = Bx^H(\omega)$ as estimate of Bx.

$$\boxed{\begin{array}{c} H=[h^1,...,h^N], \|h^j\|_2=1 \;\&\; \omega=Ax+\eta,\ \eta\sim\mathcal{SG}_\sigma, x\in\mathcal{X} \\ \Rightarrow \widehat{x}^H(\omega)=B\cdot \mathrm{argmin}_{u\in\mathcal{X}}\|H^T(Au-\omega)\|_\infty \end{array}}$$

Simple observation: *One has*

$$\mathrm{Risk}_{\|\cdot\|,\epsilon}[\widehat{x}^H|\mathcal{X}]\le \mathfrak{R}_H[\varrho(\epsilon^{-1}N)],$$
$$\varrho(s):=2\sigma\sqrt{2\ln(2s)},\ \mathfrak{R}_H[r]=\max_z\left\{\|Bz\|:\|H^TAz\|_\infty\le 2r, z\in 2\mathcal{X}_s\right\},$$

where $\mathcal{X}_s=\frac{1}{2}[\mathcal{X}-\mathcal{X}]$ *is the symmeterization of* $\mathcal{X}$*.*

Good news: $\mathfrak{R}_H[\cdot]:\mathbb{R}_+\to\mathbb{R}_+$ is concave, and $\mathfrak{R}_{[G,H]}[r]\le\min[\mathfrak{R}_G[r],\mathfrak{R}_H[r]]$

$\Rightarrow$ *whenever* $H_\ell\in\mathbb{R}^{m\times N_\ell}$, $\ell\le L$, *have* $\|\cdot\|_2$*-unit columns, it holds*

$$\mathfrak{R}_{[H_1,...,H_L]}[\varrho(\epsilon^{-1}\textstyle\sum_\ell N_\ell)]\le \mathfrak{S}\min_\ell \mathfrak{R}_{H_\ell}[\varrho(\epsilon^{-1}N_\ell)]$$
$$\mathfrak{S}=\sqrt{\frac{\ln(2\epsilon^{-1}\sum_\ell N_\ell)}{\ln(2\epsilon^{-1}\min_\ell N_\ell)}}\le\sqrt{\frac{\ln(2\sum_\ell N_\ell)}{\ln(2\min_\ell N_\ell)}}$$

$\Rightarrow$ *as far as upper bounds* $\mathfrak{R}$ *on* $(\|\cdot\|,\epsilon)$*-risks of polyhedral estimates are concerned, the "statistical price" of mimicking the best among a number of given polyhedral estimates is nearly nonexisting – all we need is to concatenate the contrast matrices of the estimates into a single matrix.*

Note: This recipe does *not* require any knowledge of risk bounds of the estimates!

$$H = [h^1, ..., h^N], \|h^j\|_2 = 1 \;\&\; \omega = Ax + \eta,\ \eta \sim \mathcal{SG}_\sigma, x \in \mathcal{X}$$
$$\Rightarrow \widehat{x}^H(\omega) = B \cdot \mathrm{argmin}_{u \in \mathcal{X}} \|H^T(Au - \omega)\|_\infty$$
$$\mathrm{Risk}_{\|\cdot\|,\epsilon}[\widehat{x}^H|\mathcal{X}] \le \mathfrak{R}_H[2\sigma\sqrt{2\ln(2\epsilon^{-1}N)}],\ \mathfrak{R}_H[r] = \max_{z \in 2\mathcal{X}_s}\{\|Bz\| : \|H^T Az\|_\infty \le 2r\}$$

- **Bad news:** *The risk bound $\mathfrak{R}_H[\cdot]$ is difficult to compute and to optimize in H.*
- **Remedy:** Under appropriate assumptions on $\mathcal{X}$ and $\|\cdot\|$, we can build efficiently computable upper bound on $\mathfrak{R}_H[\cdot]$ for which optimization in H is doable.

So far, *Remedy allowed to build, in a computationally efficient fashion, provably minimax near-optimal polyhedral estimates for the cases where*

- **A.** $\mathcal{X}_S$ *and the unit ball of* $\|\cdot\|_*$ *are spectratopes*
- **B.** $\mathcal{X} = \{x : \|Dx\|_s \le 1\}$, $\|\cdot\| = \|\cdot\|_r$, $1 \le s \le r < \infty$ *and the matrices* A, B, D *are diagonal with some restrictions on the behaviour of diagonal entries, like*

$$A = \mathrm{Diag}\{j^{-\alpha}, 1 \le j \le n\},\ B = \mathrm{Diag}\{j^{-\beta}, 1 \le j \le n\},\ D = \mathrm{Diag}\{j^{\delta}, 1 \le j \le n\}$$
$$[0 \le \alpha \le \beta,\ \delta \ge 0,\ (\beta - \alpha)r < 1]$$

Note: **B** covers several situations where linear estimates are "heavily nonoptimal."

Remark: The computation-friendly contrast design underlying the above results can be applied far beyond the situations of **A** and **B**. *What is missing, are general structural assumptions ensuring near-optimality of the estimates yielded by this design.*