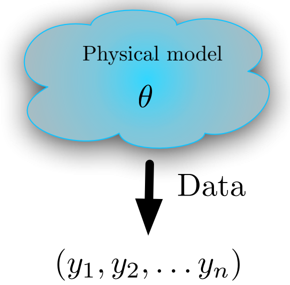
Atomic Norm Regularization for Signal and System Identification

Parikshit Shah
Collaborators: Badri Bhaskar, Gongguo Tang, Ben Recht

Inverse Problems



Challenges

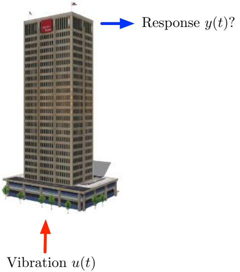
- Ill-posed, poor conditioning.
- How to find simple models?
(Today: Regularization)
- Good algorithms?
(Today: Convex optimization)

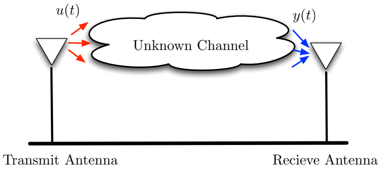
- Example: Linear inverse problem

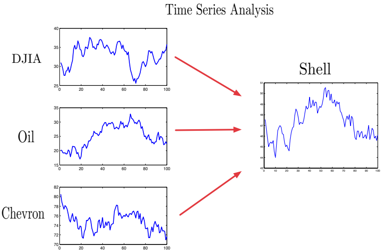
$$y = A\theta.$$

- Today: atomic norms, two specific inverse problems.

Motivating Examples







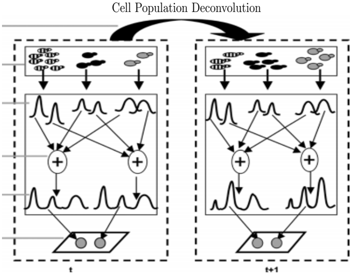
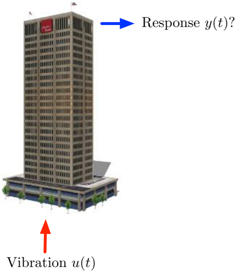
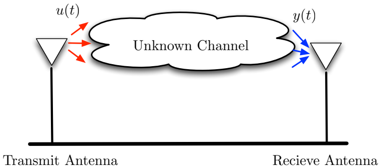


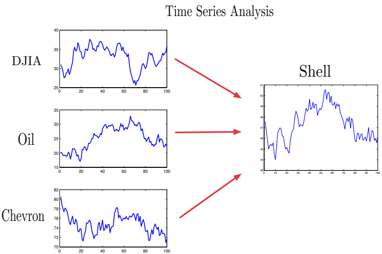
Image Courtesy: Roy et al.[1]

[1] Roy et al., A Hidden-State Markov Model for Cell Population Deconvolution, *J. Comp. Bio.*, 2006.

Motivating Examples







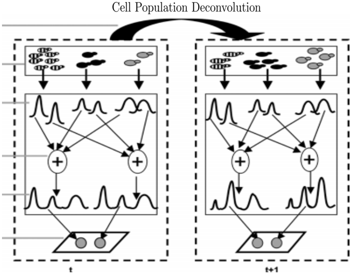
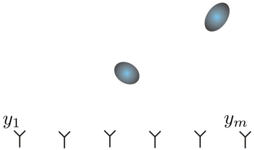
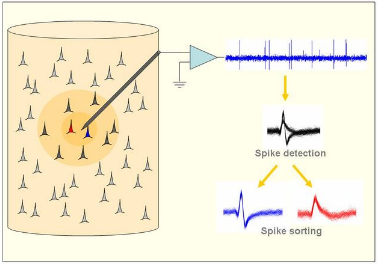


Image Courtesy: Roy et al.[1]

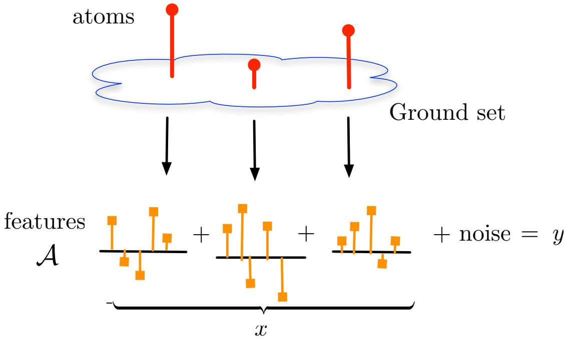
Signal identification





- Abstractly: signal identification problems
- Challenges: learn simple signal with good accuracy
- Inverse problem: massively ill-posed.

An abstract optimization setup

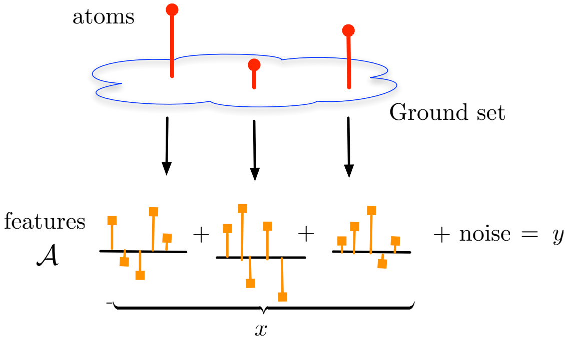


- Problem: Find x, active atoms.
- Want x to be close to y (data fidelity)

$$\underset{x}{\text{Minimize}} \qquad \|y - x\|^2$$

Want model to be simple: regularize

An abstract optimization setup

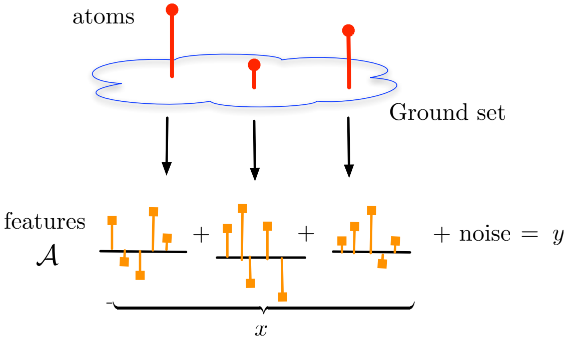


- Problem: Find x, active atoms.
- Want x to be close to y (data fidelity)

$$\underset{x}{\text{Minimize}} \qquad \|y - x\|^2$$

Want model to be simple: regularize

An abstract optimization setup



- Problem: Find x, active atoms.
- Want x to be close to y (data fidelity)

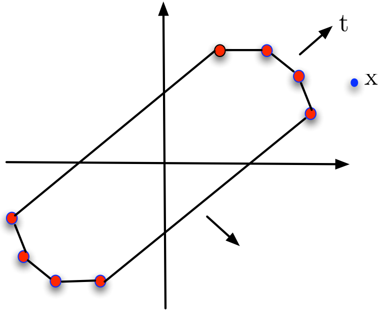
$$\underset{x}{\text{Minimize}} \qquad \|y - x\|^2 + \lambda\|x\|_{\mathcal{A}}$$

Want model to be simple: regularize

Atomic Norm

- Desire to select simple models.
- Proposal [1]: choose atomic norm.
- Ingredients: atoms, gauge norms.

$$\|x\|_{\mathcal{A}} = \inf \{t \ : \ x \in t \ \mathrm{conv}(\mathcal{A})\}$$

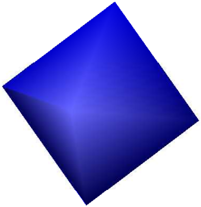


- Atomic set can be uncountably infinite.

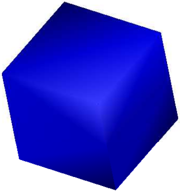
[1] V. Chandrasekaran, B. Recht, P. Parrilo, A. Willsky, *The Convex Geometry of Linear Inverse Problems.*

Special cases

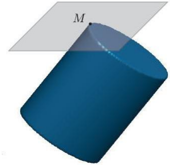
Ground set: $[n] = \{1, \dots, n\}$.
Features: $\mathcal{A} = \{\pm e_i \; : \; i \in [n]\}$.
Resulting norm: $\|c\|_1 = \sum_i |c_i|$.



Ground set: $[n] = \{1, \dots, n\}$.
$\mathcal{A} = \{(1,1,1), (1,1,-1), \dots, (-1,-1,-1)\}$
Resulting norm: $\|c\|_\infty = \max_i |c_i|$.



A more complicated example:
Matrix nuclear norm ball.



Norms can be nontrivial to optimize!

Outline

- Detour:LTI systems
- System identification
- Optimization framework
- Line spectral estimation
- Optimization framework - bounded real lemma, SDP
- Research Interests

LTI Systems

- Just linear ODEs

$$\begin{aligned} x[t+1] &= Ax[t] + Bu[t] \\ y[t] &= Cx[t] + Du[t]. \end{aligned}$$

- Equivalent representation: transfer functions

$$\text{Input } u \longrightarrow \boxed{\frac{z^2+3}{5z^3-3z+1}} \longrightarrow y \text{ Output}$$

- LTI=L (Linear) +TI (Time Invariant).
- Simplest units: simple poles $G(z) = \frac{1}{z-a}$.
- General LTI systems: $G(z) = \sum_i \frac{c_i}{z-a_i}$.

System Identification

- Unknown LTI system $G(z) = \sum_{i=1}^{s} \frac{c_i}{z-a_i}$
- Observe noisy input-output relations.
  1. Frequency response: $\mathcal{L}_k(G) = G(e^{j\omega_k})$.
  2. Impulse response: $G = \sum_{k=0}^{\infty} g_k z^{-k}$.

$$\mathcal{L}_k(G) = g_k.$$

  3. Input response: $u[k]$ known input.

$$y[k] = (g * u)[k] + w[k].$$

Abstractly:

$$y_k = \mathcal{L}_k(G) + w_k \qquad k = 1, \ldots, n.$$

- System Identification: recover G.
- Performance criterion: $\|G - \hat{G}\|_{\mathcal{H}_2}^2$.
- Assumptions: SISO, Stability radius $\rho < 1$, ...

System Identification

- Unknown LTI system $G(z) = \sum_{i=1}^{s} \frac{c_i}{z-a_i}$
- Observe noisy input-output relations.
  1. Frequency response: $\mathcal{L}_k(G) = G(e^{j\omega_k})$.
  2. Impulse response: $G = \sum_{k=0}^{\infty} g_k z^{-k}$.
  $$\mathcal{L}_k(G) = g_k.$$
  3. Input response: $u[k]$ known input.
  $$y[k] = (g * u)[k] + w[k].$$

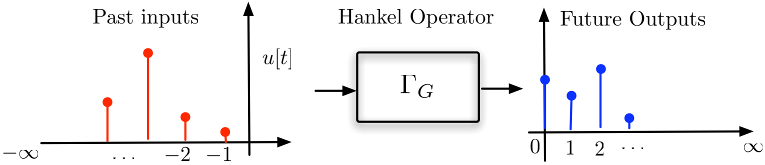
  Abstractly:
  $$y_k = \mathcal{L}_k(G) + w_k \qquad k = 1, \ldots, n.$$
- System Identification: recover G.
- Performance criterion: $\|G - \hat{G}\|_{\mathcal{H}_2}^2$.
- Assumptions: SISO, Stability radius $\rho < 1, \ldots$

Simple LTI Systems

- System ID - select simple systems.
- Simple LTI system: small McMillan degree.

$$G(z) = \sum_{i=1}^{s} \frac{c_i}{z - a_i}.$$

- Equivalent view: Hankel operators.



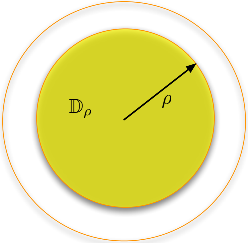
$$\Gamma_G : \ell_2(-\infty, 0) \to \ell_2[0, \infty)$$

- $\mathrm{rank}(\Gamma_G)$ = McMillan degree.
- Candidate regularizer: $\|\Gamma_G\|_*$=Hankel nuclear norm?
- Unclear how to pose.

Atomic Set for System ID

Consider the atomic set:

$$\mathcal{A} = \left\{ \pm \frac{1 - |w|^2}{z - w} \ : \ w \in \mathbb{D}_\rho \right\}.$$



Regularizer: Gauge norm generated by $\mathcal{A}$.

$$\|H\|_{\mathcal{A}} = \inf\{t \ : \ H \in t\,\mathrm{conv}(\mathcal{A})\} = \inf\left\{ \sum_{a \in \mathcal{A}} |c_a| : H = \sum_{a \in \mathcal{A}} c_a a \right\}.$$

How does this relate to $\|\Gamma_G\|_*$?

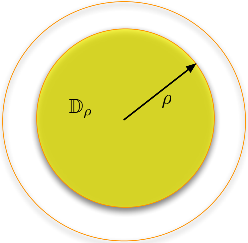
Theorem (Norm equivalence)

$$\frac{\pi}{8}\|G\|_{\mathcal{A}} \leq \|\Gamma_G\|_* \leq \|G\|_{\mathcal{A}}.$$

Atomic Set for System ID

Consider the atomic set:

$$\mathcal{A} = \left\{ \pm \frac{1 - |w|^2}{z - w} \ : \ w \in \mathbb{D}_\rho \right\}.$$



Regularizer: Gauge norm generated by $\mathcal{A}$.

$$\|H\|_{\mathcal{A}} = \inf\{t \ : \ H \in t\,\mathrm{conv}(\mathcal{A})\} = \inf\left\{ \sum_{a \in \mathcal{A}} |c_a| : H = \sum_{a \in \mathcal{A}} c_a a \right\}.$$

How does this relate to $\|\Gamma_G\|_*$?

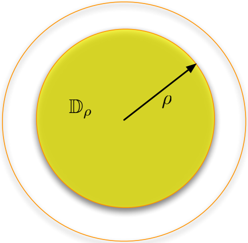
Theorem (Norm equivalence)

$$\frac{\pi}{8}\|G\|_{\mathcal{A}} \le \|\Gamma_G\|_* \le \|G\|_{\mathcal{A}}.$$

Atomic Set for System ID

Consider the atomic set:

$$\mathcal{A} = \left\{ \pm \frac{1 - |w|^2}{z - w} \ : \ w \in \mathbb{D}_\rho \right\}.$$



Regularizer: Gauge norm generated by $\mathcal{A}$.

$$\|H\|_{\mathcal{A}} = \inf\{t \ : \ H \in t\,\mathrm{conv}(\mathcal{A})\} = \inf\left\{ \sum_{a\in\mathcal{A}} |c_a| : H = \sum_{a\in\mathcal{A}} c_a a \right\}.$$

How does this relate to $\|\Gamma_G\|_*$?

Theorem (Norm equivalence)

$$\frac{\pi}{8}\|G\|_{\mathcal{A}} \leq \|\Gamma_G\|_* \leq \|G\|_{\mathcal{A}}.$$

Atomic Soft Thresholding

System ID algorithm: Solve optimization problem

$$\min_G \|y - \mathcal{L}(G)\|^2 + \lambda \|G\|_{\mathcal{A}} \qquad \text{(convex, infinite dimensional)}$$

$$\Downarrow$$

$$\begin{aligned} \min_{x,c} \quad & \|y - x\|^2 + \lambda \sum_{w \in \mathbb{D}_\rho} |c_w| \qquad \text{(infinite set of atoms)} \\ \text{subject to} \quad & x = \sum_{w \in \mathbb{D}_\rho} c_w \mathcal{L}\left(\frac{1 - |w|^2}{z - w}\right) \end{aligned}$$

$$\Downarrow$$

Discretize $\mathbb{D}_\rho$!

Atomic Soft Thresholding

System ID algorithm: Solve optimization problem

$$\min_{G} \|y - \mathcal{L}(G)\|^2 + \lambda \|G\|_{\mathcal{A}} \qquad \text{(convex, infinite dimensional)}$$

$$\Downarrow$$

$$\min_{x,c} \quad \|y - x\|^2 + \lambda \sum_{w \in \mathbb{D}_\rho} |c_w| \qquad \text{(infinite set of atoms)}$$

$$\text{subject to} \quad x = \sum_{w \in \mathbb{D}_\rho} c_w \mathcal{L}\left(\frac{1 - |w|^2}{z - w}\right)$$

$$\Downarrow$$

Discretize $\mathbb{D}_\rho$!

Atomic Soft Thresholding

System ID algorithm: Solve optimization problem

$$\min_G \|y - \mathcal{L}(G)\|^2 + \lambda \|G\|_{\mathcal{A}} \qquad \text{(convex, infinite dimensional)}$$

$$\Downarrow$$

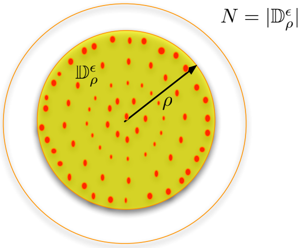
$$\min_{x,c} \quad \|y - x\|^2 + \lambda \sum_{w \in \mathbb{D}_\rho} |c_w| \qquad \text{(infinite set of atoms)}$$

$$\text{subject to} \quad x = \sum_{w \in \mathbb{D}_\rho} c_w \mathcal{L}\left(\frac{1 - |w|^2}{z - w}\right)$$

$$\Downarrow$$

Discretize $\mathbb{D}_\rho$!

Discretized Atomic Soft Thresholding (DAST)



Let

$$\Phi_{ij} = \mathcal{L}_i \left(\frac{1 - |w_j|^2}{z - w_j} \right).$$

Then formulation reduces to:

$$\min_{c \in \mathbb{R}^N} \|y - \Phi c\|^2 + \lambda \|c\|_1. \qquad (DAST)$$

Estimate: $\hat{G} := \sum_k \frac{c_k^*(1-|w_k|^2)}{z-w_k}$.

- LASSO: finite dimensional, convex, fast solvers.
- Informed discretization: no curse of dimensionality.
- Rigorous statistical guarantees.

Statistical bounds (frequency response)

Theorem (S., Bhaskar, Tang, Recht)

Let discretization level be proportional to $\frac{(1-\rho)}{n}$, *and* $\lambda = 2\sigma\sqrt{n\log\left(\frac{11\rho^2}{2(1-\rho)}\right)}$. *Let* $\hat{G}$ *be the solution obtained by DAST.*

Then there exist constants C_0, C_1, C_2 *such that:*

$$\|\hat{G} - G\|_{\mathcal{H}_2}^2 \leq \frac{C_0}{1-\rho}\left(\sigma\|\Gamma_G\|_*\sqrt{\log\left(\frac{C_1 n}{(1-\rho)}\right)}\sqrt{\frac{1}{n}} + C_2\frac{\|\Gamma_G\|_*^2}{n}\right)$$

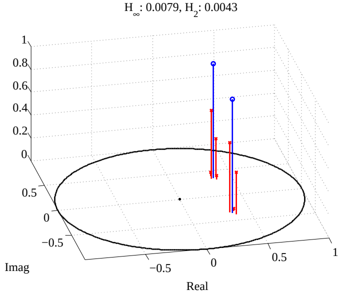
with high probability.

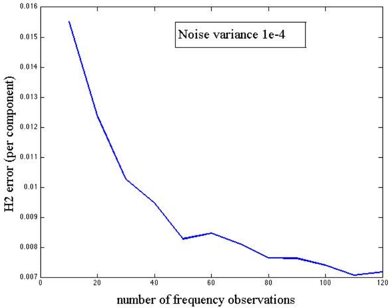
- Dominant dependence: $\|\hat{G} - G\|_{\mathcal{H}_2}^2 \leq \frac{C\|\Gamma_G\|_*}{(1-\rho)}\sqrt{\frac{\log n}{n}}$.
- Analysis: optimality conditions+operator theory+probability

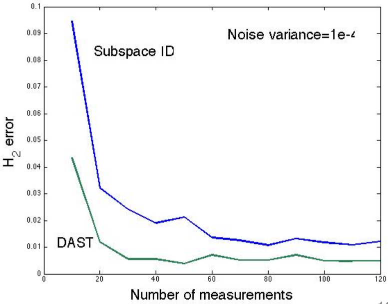
Comparison to Subspace ID

- Subspace ID-widely used algorithm.
- Manually pick the order.
- Sensitive computations.
- Not modular.
- Only asymptotic analysis available(?)

Numerical Experiments

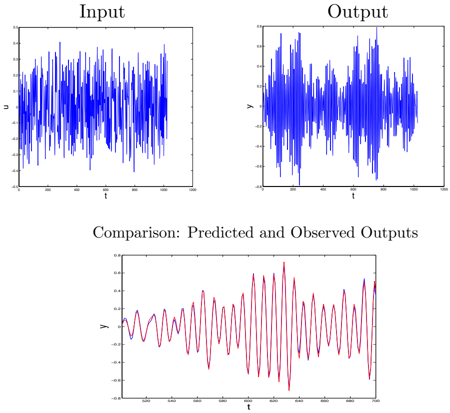






DaISy

Experimental setup involving robotic arm.
Model Applied torque $\rightarrow$ acceleration.



Relative prediction error: 10.1 percent, system order=8.
Subspace ID: produces unstable solution.

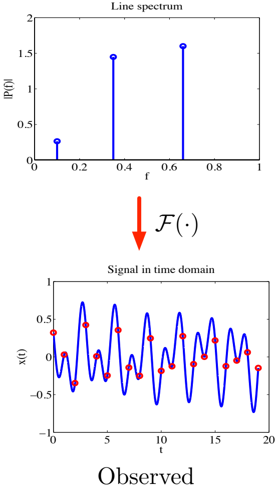
Extensions/Future Work

- MIMO case: reduce to MISO and solve:

$$\begin{aligned} \min_{x,B} \quad & \|y - x\|^2 + \lambda \sum_{w \in \mathbb{D}_\rho} \|B_w\|_2 \\ \text{subject to} \quad & x = \sum_{w \in \mathbb{D}_\rho} B_w \mathcal{L}\left(\frac{1-|w|^2}{z-w}\right) \end{aligned}$$

- Lower bounds and hints at faster rates.
- Atomic norm: finite characterization (SDP)?
- Optimal measurement ensembles?

Line Spectral Estimation

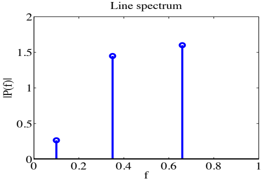


- Problem: recover spectrum from samples.
- Robustness Issues: missing samples, noise.

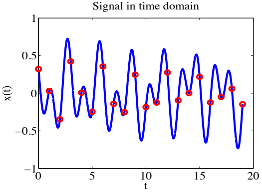
Line spectral estimation

$$\mu(f) = \sum_{l=1}^{k} c_l \delta(f - f_l), \quad f_l \in [0, 1]$$

$$y = \overbrace{\sum_{l=1}^{k} c_l \underbrace{\frac{1}{\sqrt{n}} \begin{bmatrix} 1 \\ e^{i2\pi f_l} \\ e^{i2\pi 2 f_l} \\ \vdots \\ e^{i2\pi (n-1) f_l} \end{bmatrix}}_{a(f_l)}}^{x} + w$$



$\mathcal{F}(\cdot)$

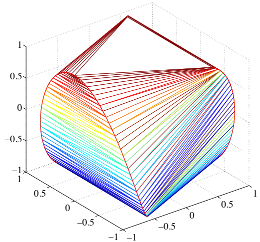


Observed

- Denoising: Given y, recover x.
- Measure recovery: Given y, recover c_l, f_l.
- Prior art: Prony, Cadzow, [Candes, Fernandez-Granda 2012].

Atomic norm denoising

Atomic set $\mathcal{A}: e^{i\phi} a(f) = e^{i\phi} \begin{bmatrix} 1 \\ e^{i2\pi f} \\ e^{i2\pi 2f} \\ \vdots \\ e^{i2\pi(n-1)f} \end{bmatrix}$



The atomic norm is:

$$\|x\|_{\mathcal{A}} = \inf\{t : x \in t\,\mathrm{conv}(\mathcal{A})\} = \inf\{\sum_l |c_l|, x = \sum_l c_l a(f_l)\}.$$

Denoising algorithm: Solve

$$\mathrm{AST}: \qquad \underset{z}{\text{minimize}}\ \frac{1}{2}\|y - z\|_2^2 + \tau\|z\|_{\mathcal{A}}$$

- Analyzed in [Bhaskar, Tang, Recht 2012].

Atomic norm and Bounded Real Lemma

How does one solve AST?

Exact Refomulation as a semidefinite program!

$$\begin{array}{ll} \underset{z,t,u}{\text{minimize}} & \frac{1}{2}\|y - z\|_2^2 + \frac{\tau}{2}(t + nu_0) \\ \text{subject to} & \\ & \begin{bmatrix} \mathrm{Toep}\,(u) & z \\ z^* & t \end{bmatrix} \succcurlyeq 0 \end{array}$$

Three ingredients:

1. Dual formulation: $\|x\|_{\mathcal{A}} = \sup_{\lambda : \|\lambda\|_{\mathcal{A}}^* \leq 1} \langle x, \lambda \rangle$
2. Dual norm characterization: $\|\lambda\|_{\mathcal{A}}^* = \sup_{f : f \in [0,1]} \frac{1}{\sqrt{n}} \underbrace{\left| \sum_{l=0}^{n-1} \lambda_l e^{-i2\pi lf} \right|}_{\Lambda(f)}$
3. Bounded Real Lemma: A trigonometric polynomial $|\Lambda(f)| \leq \sqrt{n}, \forall f \in [0,1]$ if and only if $\exists Q \succeq 0$ such that

$$\mathrm{Toep}^*(Q) = ne_1, \begin{bmatrix} Q & \lambda \\ \lambda^* & 1 \end{bmatrix} \succeq 0.$$

Atomic norm and Bounded Real Lemma

How does one solve AST?

Exact Refomulation as a semidefinite program!

$$\begin{array}{ll} \underset{z,t,u}{\text{minimize}} & \frac{1}{2}\|y - z\|_2^2 + \frac{\tau}{2}(t + nu_0) \\ \text{subject to} & \\ & \begin{bmatrix} \text{Toep}\,(u) & z \\ z^* & t \end{bmatrix} \succcurlyeq 0 \end{array}$$

Three ingredients:

1. Dual formulation: $\|x\|_{\mathcal{A}} = \sup_{\lambda : \|\lambda\|_{\mathcal{A}}^* \leq 1} \langle x, \lambda \rangle$
2. Dual norm characterization: $\|\lambda\|_{\mathcal{A}}^* = \sup_{f : f \in [0,1]} \frac{1}{\sqrt{n}} \underbrace{\left| \sum_{l=0}^{n-1} \lambda_l e^{-i2\pi l f} \right|}_{\Lambda(f)}$
3. Bounded Real Lemma: A trigonometric polynomial $|\Lambda(f)| \leq \sqrt{n}, \forall f \in [0, 1]$ if and only if $\exists Q \succeq 0$ such that

$$\text{Toep}^*(Q) = ne_1, \begin{bmatrix} Q & \lambda \\ \lambda^* & 1 \end{bmatrix} \succeq 0.$$

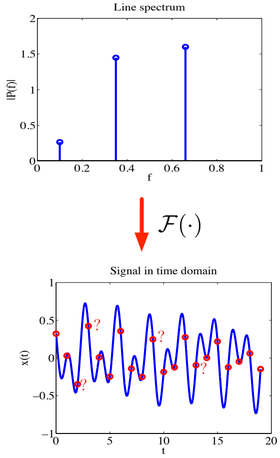
Missing samples

Can we fill in the missing samples?

True samples: $x^* = (x_1^*, \ldots, x_m^*)$
Observed samples: $x_T^* = (x_{i_1}^*, \ldots x_{i_m}^*)$

Recovery algorithm:

$$\begin{array}{ll} \underset{z}{\text{minimize}} & \|z\|_{\mathcal{A}} \\ \text{subject to} & z_T = x_T^*. \end{array}$$



Observed

Recovery Guarantee

Theorem (Tang, Bhaskar, S., Recht 2012)

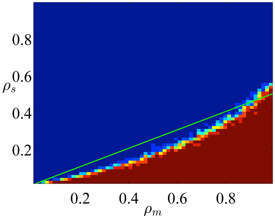
Suppose we observe (random) time samples of x^* in interval $\{0, \ldots, n\}$, and sign(c_k) i.i.d. on $\mathbb{C}$. If $\Delta_f \geq \frac{4}{n}$, then

$$C \max \left\{ \log^2 \frac{n}{\delta}, k \log \frac{k}{\delta} \log \frac{n}{\delta} \right\},$$

measurements are sufficient to guarantee that with probability at least $1 - \delta$, $x^\star$ is the unique optimizer.

Alternative viewpoint:
Nonuniform sub-Nyquist sampling sufficient!

Connections to compressive sensing.

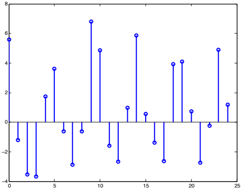


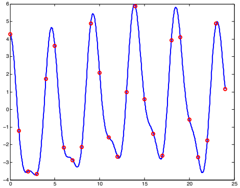
Deconvolution

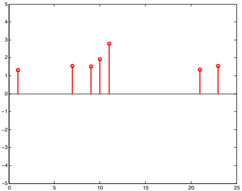
Suppose we are given

$$y = \sum_{l=1}^{k} c_l \frac{1}{\sqrt{n}} \begin{bmatrix} 1 \\ e^{i2\pi f_l} \\ e^{i2\pi 2 f_l} \\ \vdots \\ e^{i2\pi (n-1) f_l} \end{bmatrix} + s,$$

where s is a S-sparse vector.







$$y \quad = \quad x \quad + \quad s$$

Given y, how to deconvolve into x and s?

Application: Poisson noise.

Deconvolution

Solve:

$$\text{ADC}: \quad \underset{x,s}{\text{minimize}} \quad \|x\|_{\mathcal{A}} + \|s\|_1 \qquad \text{subject to} \quad y = x + s.$$

Theorem (Tang, Bhaskar, S., Recht 2012)

Suppose $y = x + s \in \mathbb{C}^n$ and

1. $\Delta f \geq \frac{4}{n}$, *(Frequency separation)*
2. $\sqrt{n} > C \max\{k, S\}$ *(Sufficient samples)*

Then ADC recovers the true solution.

Proof Idea

Analyze optimality conditions, exhibit explicit dual certificate, involves polynomial interpolation theory.

Deconvolution

Solve:

$$\text{ADC}: \quad \underset{x,s}{\text{minimize}} \quad \|x\|_{\mathcal{A}} + \|s\|_1$$
$$\text{subject to} \quad y = x + s.$$

Theorem (Tang, Bhaskar, S., Recht 2012)

Suppose $y = x + s \in \mathbb{C}^n$ and

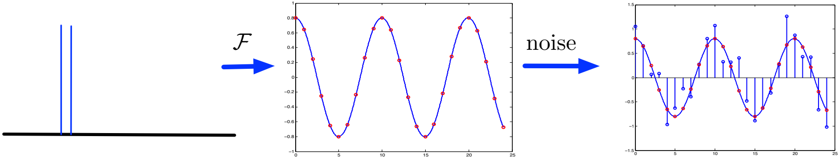
1. $\Delta f \geq \frac{4}{n}$, *(Frequency separation)*
2. $\sqrt{n} > C \max\{k, S\}$ *(Sufficient samples)*

Then ADC recovers the true solution.

Proof Idea

Analyze optimality conditions, exhibit explicit dual certificate, involves polynomial interpolation theory.

Ongoing work: limits of robustness and resolution



- Two issues: resolvability and robustness to noise.

How small is the denoising error?

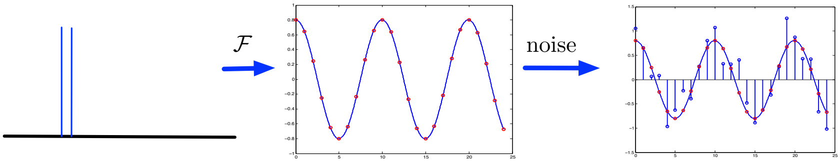
Conjecture (Optimal Statistical Rates)

If k is small and the frequencies are well-separated, then w.h.p. solution to AST

$$\frac{1}{n}\|\hat{x} - x\|_2^2 \leq O\left(\frac{k \log(n)}{n}\sigma^2\right).$$

- Approximate support recovery

Ongoing work: limits of robustness and resolution



- Two issues: resolvability and robustness to noise.

How small is the denoising error?

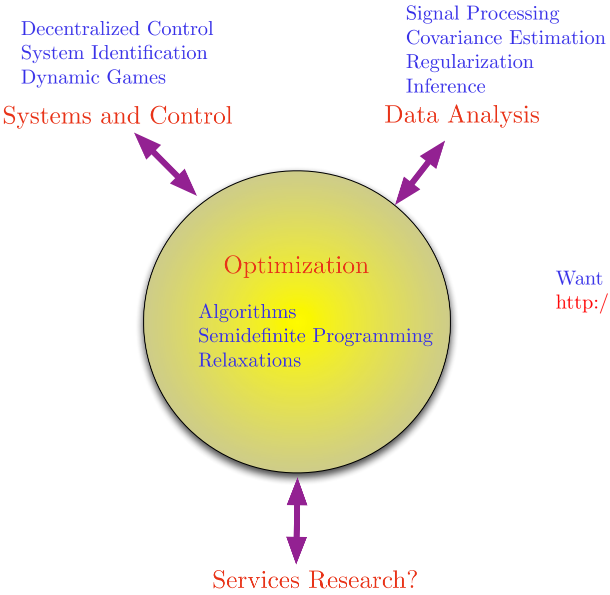
Conjecture (Optimal Statistical Rates)

If k is small and the frequencies are well-separated, then w.h.p. solution to AST

$$\frac{1}{n}\|\hat{x} - x\|_2^2 \le O\left(\frac{k\log(n)}{n}\sigma^2\right).$$

- Approximate support recovery

Research



Want to know more?
http://people.lids.mit.edu/pari/

Line spectral estimation