Recovery and denoising with simultaneous structures

Maryam Fazel

University of Washington

Joint work with:
Samet Oymak (Berkeley), Amin Jalali (UW),
Babak Hassibi (Caltech), Yonina Eldar (Technion)

SPARS 2015, Cambridge University, July 7, 2015

Outline

- simultaneous structures: where and why?
- review: single structure case
  compressed sensing, low-rank recovery,. . .
- fundamental limitation of combining convex penalties, for
  - arbitrary norms
  - a variety of measurements, beyond Gaussian
- similar result for the problem of ‘denoising’
- what next?

Low-dimensional structures

classic examples:

- sparse vectors (e.g., compressed sensing) ℓ_1 norm
- group-sparse vectors (group LASSO) $\ell_{1,2}$ norm
- low-rank matrices (collaborative filtering, phase retrieval,. . .)   nuclear (trace) norm
- sparse *plus* low-rank matrices, $\mathbf{X} = \mathbf{L} + \mathbf{S}$ (PCA with outliers, graphical models with hidden variables)
  ℓ_1 plus nuclear norm

Low-dimensional structures

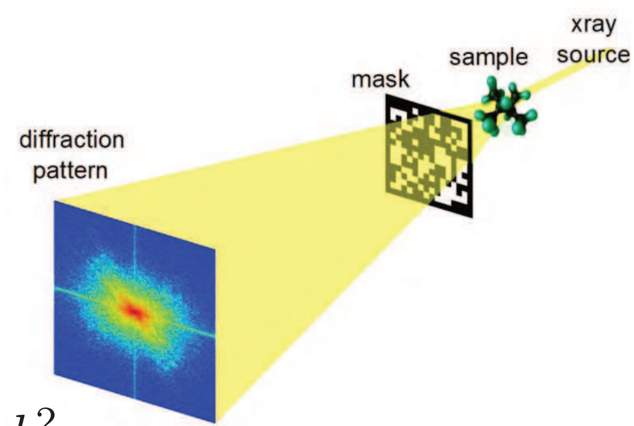
multiple, simultaneous structures

- simultaneously sparse *and* low-rank matrices (sparse phase retrieval, sparse PCA, quadratic compressed sensing,. . .) ℓ_1 and nuclear norms
- tensors with low Tucker rank
  nuclear norms of unfolded matrices
- simultaneously sparse and piece-wise constant vectors (e.g., 'fused lasso')
  ℓ_1 norm and total-variation norm, $\|\mathbf{x}\|_{TV} = \sum_{i=1}^{n-1} |\mathbf{x}_{i+1} - \mathbf{x}_i|$

Sparse and low-rank matrices: an application

phase retrieval, a classic signal processing/optics problem

recover signal $\mathbf{x}_0$ from linear *phaseless* measurements,



$$|\mathbf{a}_i^T \mathbf{x}_0| = b_i, \quad i = 1, \ldots, m$$

reformulate as: find $\mathbf{X} = \mathbf{x}_0 \mathbf{x}_0^T$ s.t. $\langle \mathbf{a}_i \mathbf{a}_i^T, \mathbf{X} \rangle = b_i^2$

i.e., $\mathbf{X} \succeq 0$, $\mathrm{rank}(\mathbf{X}) = 1$, $\mathcal{A}(\mathbf{X}) = b'$ [Candes,Eldar,Strohmer,Voroninski'11]

signal $\mathbf{x}_0$ can also be **sparse**. then, $\mathbf{X}$ is rank-1 and (block-)sparse.

other applications (for sparse and low-rank matrices):

- sparse PCA [d'Aspremont et al'08,. . .]
  - find approximate eigenvectors of $\mathbf{X}$ that are sparse, e.g., $\mathbf{X} \approx \mathbf{x}\mathbf{x}^T$ with $\mathbf{x}$ k-sparse
- cluster detection [Richard,Savalle,Vayatis'12]
  - ideal cluster adjacency matrix is low-rank & sparse

Recovery of structured models

unknown structured model $\mathbf{x}_0 \in \mathbf{R}^n$

- recovery from **compressed measurements:** $\mathcal{A}(\mathbf{x}_0) = \mathbf{y}$
  linear $\mathcal{A} : \mathbf{R}^n \to \mathbf{R}^m$, $m \ll n$. can write as $\mathbf{A}\mathbf{x} = y$ with $\mathbf{A} \in \mathbf{R}^{m \times n}$
- **denoising:** $\mathcal{A}$ is identity; $\mathbf{y} = \mathbf{x}_0 + \mathbf{z}$, noise $\mathbf{z}$ is i.i.d
- LASSO: $\mathbf{y} = \mathcal{A}(\mathbf{x}_0) + \mathbf{z}$

goal: given $\mathcal{A}$ and $\mathbf{y} \in \mathbf{R}^m$, find $\mathbf{x}_0$.

- how many measurements m suffice? (sample complexity)
- how does mean-squared error behave with noise level?

Example: Sparse vectors and $\|\mathbf{x}\|_1$

$\mathcal{A} : \mathbf{R}^n \to \mathbf{R}^m$, suppose $\mathcal{A}$ is Gaussian. $\mathbf{x}_0$ is k-sparse.

non-convex program:

$$\begin{array}{ll} \text{minimize} & \|\mathbf{x}\|_0 \\ \text{subject to} & \mathcal{A}(\mathbf{x}) = \mathcal{A}(\mathbf{x}_0) \end{array}$$

needs $\mathcal{O}(k)$ observations to exactly recover $\mathbf{x}_0$ with high probability*

convex program:

$$\begin{array}{ll} \text{minimize} & \|\mathbf{x}\|_1 \\ \text{subject to} & \mathcal{A}(\mathbf{x}) = \mathcal{A}(\mathbf{x}_0) \end{array}$$

needs $\mathcal{O}(k \log n)$ observations for exact recovery w.h.p.

$*$ means: there exists constant c s.t. $\mathbf{x}_0$ is found with probability $> 1 - \exp(-cm)$

[Candes,Romberg,Tao'04; Donoho'04; Tropp'04; Fuchs'04; . . .]

Example: Low-rank matrices and $\|\mathbf{X}\|_*$

$\mathcal{A} : \mathbf{R}^{n\times n} \to \mathbf{R}^m$, suppose $\mathcal{A}$ is Gaussian. $\mathbf{X}_0$ is rank r.

non-convex program:

$$\begin{array}{ll} \text{minimize} & \mathrm{rank}(\mathbf{X}) \\ \text{subject to} & \mathcal{A}(\mathbf{X}) = \mathcal{A}(\mathbf{X}_0) \end{array}$$

needs $\mathcal{O}(nr)$ observations to exactly recover $\mathbf{X}_0$ w.h.p.

convex program:

$$\begin{array}{ll} \text{minimize} & \|\mathbf{X}\|_* \\ \text{subject to} & \mathcal{A}(\mathbf{X}) = \mathcal{A}(\mathbf{X}_0) \end{array}$$

also needs $\mathcal{O}(nr)$ observations for exact recovery w.h.p.

[Recht,Fazel,Parrilo'07; Candes,Recht'08; Candes,Plan'09; Negahban,Wainwright'09,. . .]

also true for other classic examples:

- sparse vectors (e.g., compressed sensing) ℓ_1 norm
- group-sparse vectors (group LASSO) $\ell_{1,2}$ norm
- low-rank matrices (collaborative filtering, phase retrieval,. . .) nuclear (trace) norm
- sparse *plus* low-rank matrices, $\mathbf{X} = \mathbf{L} + \mathbf{S}$ (compressive PCA, . . .) ℓ_1 plus nuclear norm

Simultaneously structured $\mathbf{x}_0$

- object $\mathbf{x}_0$ has *several* structures, each with a structure-promoting norm
- additional structures reduce degrees of freedom

consider class of convex programs

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) = h(\|\mathbf{x}\|_{(1)}, \ldots, \|\mathbf{x}\|_{(S)}) \\ \text{subject to} & \mathcal{A}(\mathbf{x}) = \mathcal{A}(\mathbf{x}_0) \end{array}$$

where $h : \mathbf{R}_+^S \to \mathbf{R}_+$ is convex and non-decreasing in each argument

examples:

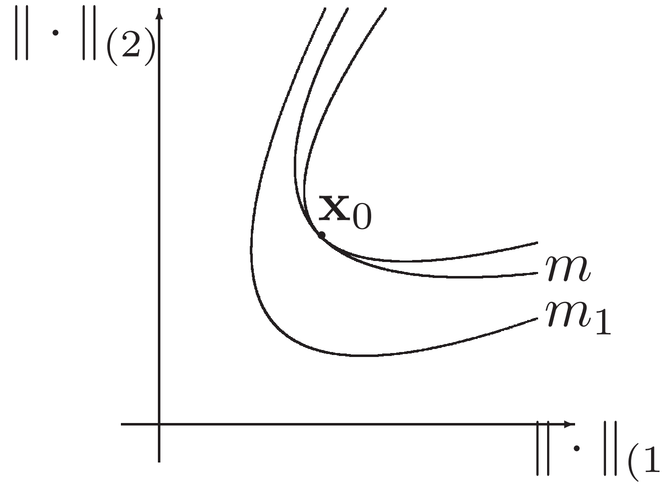
$$f(\mathbf{x}) = \sum_{i=1}^{S} \lambda_i \|\mathbf{x}\|_{(i)}, \qquad f(\mathbf{x}) = \max_{i=1,\ldots,S} \alpha_i \|\mathbf{x}\|_{(i)}$$

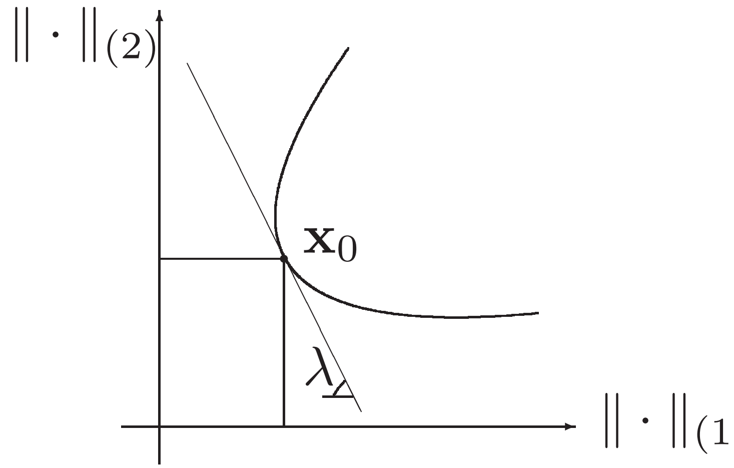
$\lambda_i, \alpha_i > 0$ are parameters

Pareto optimal front

pick m. consider set of norm values achieved by $\{\mathbf{x} \mid \mathcal{A}(\mathbf{x}) = \mathcal{A}(\mathbf{x}_0)\}$ and fill the upper-right points to get the Pareto optimal set for each m. observe

- if we have $m_1 < m$ measurements, $\mathbf{x}_0$ doesn't correspond to Pareto optimal front
  - cannot be recovered by minimizing *any* combination of norms
- need at least m measurements for $\mathbf{x}_0$ to be recoverable





Some results

- a limitation for combining convex penalties: simpler proof
- holds true for a variety of measurements $\mathcal{A}$:
  - Gaussian iid entries
  - independent subgaussian rows
  - sampling operator
    (e.g., sampled rows of identity as in 'completion' problems, or sampled rows of Fourier matrix)
  - quadratic (or rank-1) measurements: $\langle \mathbf{a}_i \mathbf{a}_i^T, \mathbf{X} \rangle = b_i^2$
- special case: sparse and low-rank matrix

[Oymak et al. '12,'15]

Recovery failure: sufficient condition

suppose $\mathbf{x}_0$ has structures $i = 1, \ldots, S$. when does program

$$\text{minimize} \;\; f(\mathbf{x}) = h(\|\mathbf{x}\|_{(1)}, \ldots, \|\mathbf{x}\|_{(S)}) \quad \text{subject to} \;\; \mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x}_0$$

fail to give $\mathbf{x}_0$ as its solution?

theorem. if

$$\inf_{\mathbf{g} \in \partial f(\mathbf{x}_0)} |\bar{\mathbf{g}}^T \bar{\mathbf{x}}_0| \quad > \quad \frac{\|\mathbf{A}\bar{\mathbf{x}}_0\|_2}{\sigma_{\min}(\mathbf{A})},$$

where $\bar{\mathbf{x}}_0 = \frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2}$, $\bar{\mathbf{g}} = \frac{\mathbf{g}}{\|\mathbf{g}\|_2}$, then $\mathbf{x}_0$ is **not** a minimizer and recovery fails.

($\partial f(\mathbf{x}_0)$ is the set of subgradients of f at $\mathbf{x}_0$)

Recovery failure: sufficient condition

theorem. if

$$\inf_{\mathbf{g} \in \partial f(\mathbf{x}_0)} |\bar{\mathbf{g}}^T \bar{\mathbf{x}}_0| \;\; > \;\; \frac{\|\mathbf{A}\bar{\mathbf{x}}_0\|_2}{\sigma_{\min}(\mathbf{A})},$$

then $\mathbf{x}_0$ is **not** a minimizer and recovery fails.

- LHS depends only on f and $\bar{\mathbf{x}}_0$
- cannot be made too small, as subgradients are 'aligned' with $\bar{\mathbf{x}}_0$ (we bound this with a geometric quantity)
- RHS depends only on $\mathbf{A}$ and $\bar{\mathbf{x}}_0$
- for many random ensembles, RHS $\gtrsim \sqrt{\frac{m}{n}}$

Recovery failure: sufficient condition

theorem. if

$$\inf_{\mathbf{g} \in \partial f(\mathbf{x}_0)} |\bar{\mathbf{g}}^T \bar{\mathbf{x}}_0| \quad > \quad \frac{\|\mathbf{A}\bar{\mathbf{x}}_0\|_2}{\sigma_{\min}(\mathbf{A})},$$

then $\mathbf{x}_0$ is **not** a minimizer and recovery fails.

- LHS depends only on $f(\cdot)$ and $\bar{\mathbf{x}}_0$
- cannot be made too small, as subgradients are 'aligned' with $\bar{\mathbf{x}}_0$
- RHS depends only on $\mathbf{A}$ and $\bar{\mathbf{x}}_0$
- for many random ensembles, RHS $\gtrsim \sqrt{\frac{m}{n}}$

Recovery failure: sufficient condition

theorem. if

$$\inf_{\mathbf{g} \in \partial f(\mathbf{x}_0)} |\bar{\mathbf{g}}^T \bar{\mathbf{x}}_0| \;>\; \frac{\|\mathbf{A}\bar{\mathbf{x}}_0\|_2}{\sigma_{\min}(\mathbf{A})}$$

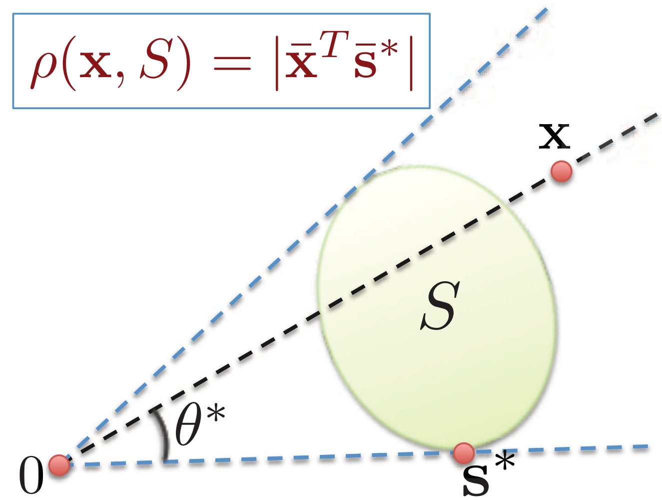
then $\mathbf{x}_0$ is **not** a minimizer and recovery fails.

- LHS depends only on f and $\bar{\mathbf{x}}_0$
- cannot be made too small, as subgradients are 'aligned' with $\bar{\mathbf{x}}_0$
- RHS depends only on $\mathbf{A}$ and $\bar{\mathbf{x}}_0$
- for many random ensembles, RHS $\approx \sqrt{\frac{m}{n}}$

Bound the left-hand side

def.: *correlation* between $\mathbf{x}_0$ and set S (largest angle)

$$\rho(\mathbf{x}, S) = \inf_{0 \neq \mathbf{s} \in S} |\bar{\mathbf{x}}^T \bar{\mathbf{s}}|$$



if set S is subdiff. of norm i:

$$\rho(\mathbf{x}_0, \partial\|\mathbf{x}_0\|_{(i)}) = \frac{\|\bar{\mathbf{x}}_0\|_{(i)}}{\sup_{g \in \partial\|\mathbf{x}_0\|_{(i)}} \|g\|_2} \geq \frac{\|\bar{\mathbf{x}}_0\|_{(i)}}{L_i} := \kappa_i$$

where L_i is the norm's Lipschitz constant. now lower bound the LHS,

$$\inf_{\mathbf{g} \in \partial f(\mathbf{x}_0)} |\bar{\mathbf{g}}^T \bar{\mathbf{x}}_0| \;\geq\; \kappa_{\min} = \min_i \kappa_i.$$

(see also [Mu,Huang,Wright,Goldfarb'13]) proof:

- from convex analysis:
  any subgradient of $f = h(\|\mathbf{x}\|_{(1)}, \ldots, \|\mathbf{x}\|_{(S)})$ can be written as $\mathbf{g} = \sum_i w_i \mathbf{g}_i$ with $w_i \geq 0$, where $\mathbf{g}_i$ is a subgradient of $\|\mathbf{x}\|_{(i)}$

- $\mathbf{g}^T \bar{\mathbf{x}}_0 = \sum_i w_i \|\bar{\mathbf{x}}_0\|_{(i)}$ (since $\mathbf{g}_i^T \bar{\mathbf{x}}_0 = \|\bar{\mathbf{x}}_0\|_{(i)}$)

- $\|\mathbf{g}\|_2 \leq \sum_i w_i \|\mathbf{g}_i\|_2 \leq \sum_i w_i L_i$, so

$$\inf_{\mathbf{g} \in \partial f(\mathbf{x}_0)} |\bar{\mathbf{g}}^T \bar{\mathbf{x}}_0| \geq \frac{\sum_i w_i \|\bar{\mathbf{x}}_0\|_{(i)}}{\sum_i w_i L_i} \geq \min_i \frac{w_i \|\bar{\mathbf{x}}_0\|_{(i)}}{w_i L_i} = \kappa_{\min}.$$

Bound the right-hand side (via random matrix theory)

random vector $\mathbf{x} \in \mathbf{R}^n$ is subgaussian, if marginals $\mathbf{x}^T\mathbf{v}$ are subgaussian random variables for all $\mathbf{v} \in \mathbf{R}^n$

lemma [subgaussian measurements] if $\mathbf{A}$ has

- i.i.d zero-mean isotropic subgaussian rows, or
- i.i.d zero-mean, unit-variance subgaussian entries

there exists constant c_1 such that whenever $m \le c_1 n$, w.h.p. we have

$$\frac{\|\mathbf{A}\bar{\mathbf{x}}_0\|_2^2}{\sigma_{\min}^2(\mathbf{A})} \le \frac{2m}{n}$$

[see review by Vershynin,'14]

Bound the left-hand side

lemma [sampling] sample rows of $\mathbf{A}$ uniformly from the identity matrix, discard duplicate rows. then w.h.p.,

$$\frac{\|\mathbf{A}\bar{\mathbf{x}}_0\|_2^2}{\sigma_{\min}^2(\mathbf{A})} \leq \frac{2m}{n}$$

several other measurements give similar bounds, e.g., $\langle \mathbf{a}_i \mathbf{a}_i^T, \mathbf{X} \rangle = b_i^2$ (also studied in [Li, Voroninski '12])

Recovery failure

putting bounds together:

theorem. $\mathbf{x}_0$ will not be a minimizer of the recovery program w.h.p., if

$$m \;\leq\; c\, n\kappa_{\min}^2$$

for all measurement types mentioned.

examples:

model	$f(\cdot)$	L	$\|\bar{\mathbf{x}}_0\| \leq$	m at least
k sparse vector	$\|\cdot\|_1$	$\sqrt{n}$	$\sqrt{k}$	k
k column-sparse matrix	$\|\cdot\|_{1,2}$	$\sqrt{d}$	$\sqrt{k}$	kd
rank r matrix	$\|\cdot\|_\star$	$\sqrt{d}$	$\sqrt{r}$	rd
sparse & Low-rank matrix	$h(\|\cdot\|_\star, \|\cdot\|_1)$	—	—	$\min\{k^2, rd\}$

last three lines: $d \times d$ matrix with $k \times k$ nonzero block, rank r, and $n = d^2$

Sparse and low-rank case

a gap. a nonconvex problem can recover the model from few measurements (on order of the degrees of freedom), while combined convex penalties requires much more measurements (suppose $\mathcal{A}$ is Gaussian).

m

$\mathcal{O}\left(nr\right)$

$\mathbf{X}_0$ is *not* a minimizer of

$$\begin{array}{ll} \text{min.} & \|\mathbf{X}\|_* + \lambda \|\mathbf{X}\|_{1,2} \\ \text{s.t.} & \mathcal{A}(\mathbf{X}) = \mathcal{A}(\mathbf{X}_0) \\ & \mathbf{X} \succeq 0 \end{array}$$

gap

$\mathbf{X}_0$ is the *unique* minimizer of

$\mathcal{O}\left(k \log \frac{n}{k} + kr\right)$

$$\begin{array}{ll} \text{min.} & \operatorname{rank}(\mathbf{X}) + \lambda \|\mathbf{X}\|_{0,2} + \lambda \|\mathbf{X}^T\|_{0,2} \\ \text{s.t.} & \mathcal{A}(\mathbf{X}) = \mathcal{A}(\mathbf{X}_0) \end{array}$$

$r(2k-r)$ "degrees of freedom"

Numerical experiments

grayscale shows probability of success over 100 runs for each case. recovery using $f(\mathbf{X}) = \mathrm{Tr}(\mathbf{X}) + \lambda \|\mathbf{X}\|_1$. $\mathbf{X}_0$ is PSD, rank 1, $k = 8$.

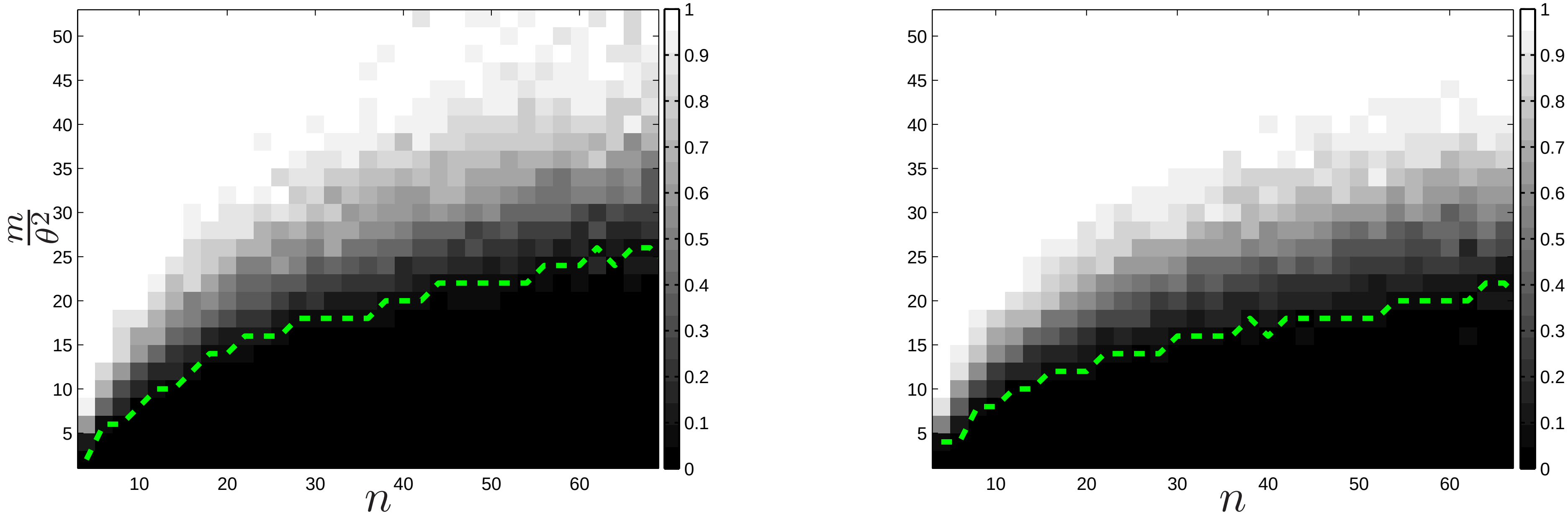


Figure 1: $\lambda = 0.2$ (left) and $\lambda = 0.35$ (right).

A related problem: Denoising

this bottleneck also appears in another problem:

suppose $\mathbf{x}_0$ has S structures; estimate $\mathbf{x}_0 \in \mathbf{R}^n$ given $\mathbf{y} = \mathbf{x}_0 + \mathbf{z}$, where $\mathbf{z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$. use:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \{ \|\mathbf{y} - \mathbf{x}\|_2^2 + \sum_{i=1}^{S} \lambda_i \|\mathbf{x}\|_{(i)} \}$$

(aka proximal operator of function $\sum_{i=1}^{S} \lambda_i \|\mathbf{x}\|_{(i)}$)

def. the MSE risk of above estimator at $\mathbf{x}_0$ is

$$\eta(\lambda_1, \ldots \lambda_S) = \max_{\gamma > 0} \mathbf{E}[\|\hat{\mathbf{x}}(\gamma \mathbf{x}_0 + \mathbf{z}) - \gamma \mathbf{x}_0\|_2^2$$

how low can MSE risk get?

can show: performance is order-wise the same as using the best single norm

similar statements for the case $y = \mathcal{A}(\mathbf{x}) + \mathbf{z}$

Summary

- simultaneously structured models: weighted sum of norms is often used in applications, lacked performance theory

- result: combined convex penalty displays a fundamental gap, both for recovery sample complexity and denoising error

- lower bound holds for various measurements, e.g., sampling (matrix or tensor completion), quadratic measurements (phase retrieval, sparsePCA)

- tight *upper* bounds on m can be obtained for the Gaussian case, for lin comb of norms with λ_i's tuned optimally; differs from lower bounds by a log factor [Oymak et al, 2015]

Discussion: what next?

is the situation all gloomy. . . ?

- find better penalty/regularizer:
  - can we directly define atoms and take convex hulls to find the atomic norm? [Chandrasekaran et al'10]
    seems intractable for sparse and low-rank case, but may help in other problems
  - convex relaxation hierarchies for the atomic norm
  - some improvements (though not orderwise) on a case-by-case basis:
    * tensors with low Tucker rank [Mu, Huang, Wright, Goldfarb '13]
    * a relaxation for sparse and low-rank [Richard, ']

- find more suitable measurements schemes (e.g., sequential measurements [Bahmani, Romberg '15])

References

"Simultaneously Structured Models with Application to Sparse and Low-rank Matrices", S. Oymak, A. Jalali, M. Fazel, Y. Eldar, B. Hassibi, *IEEE Trans. Info. Theory*, vol 61 (5), 2015. arXiv:1212.3753.

"Noisy Estimation of Simultaneously Structured Models: Limitation of Convex Relaxation", S. Oymak, A. Jalali, M. Fazel, B. Hassibi. CDC 2013.