Generic Chaining Meets (Non)convex Optimization

Mahdi Soltanolkotabi

June 23, 2016
Workshop on Chaining with Applications to Computer Science
Harvard University

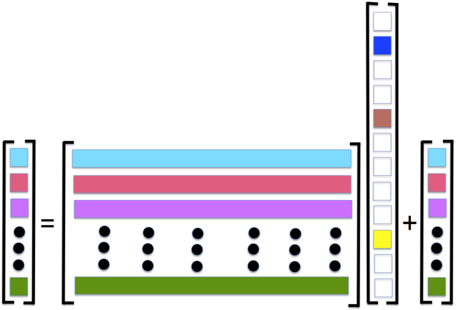
Ming Hsieh Department of Electrical Engineering

USC University of Southern California

Collaborators: Samet Oymak and Ben Recht

Xiaodong Li and Emmanuel Candes

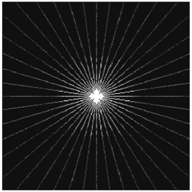
Linear inverse problems



$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$$

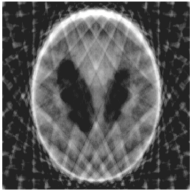
Compressed Sensing

Fourier transform



highly subsampled

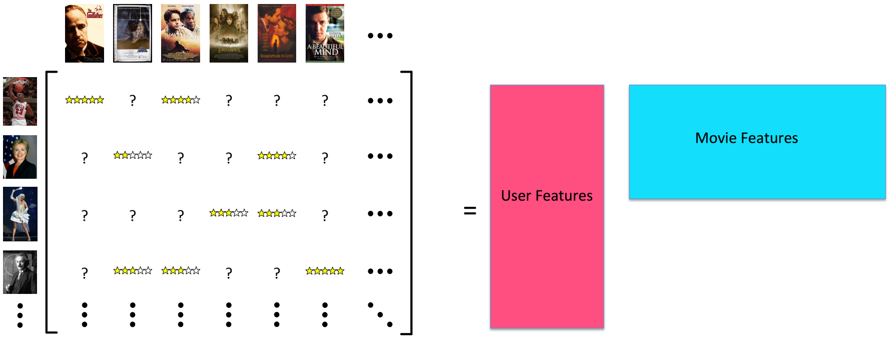
classical reconstruction



compressed sensing
reconstruction

[Photo credit Candes, Romberg, Tao 2006]
linear measurements from a structured signal.

Recommender systems



The “power” of convex programing

Exciting research over the last decade demonstrating the effectiveness of convex programming/greedy algorithms.

The “power” of convex programing

Exciting research over the last decade demonstrating the effectiveness of convex programming/greedy algorithms.

Idealogy

“when life gives you lemons, convexify”

The "power" of convex programing

Exciting research over the last decade demonstrating the effectiveness of convex programming/greedy algorithms.

Idealogy

"when life gives you lemons, convexify"

- Sparse use ℓ_1 norm, Low-rank use nuclear norm, atomic norms, etc.

The “power” of convex programing

Exciting research over the last decade demonstrating the effectiveness of convex programming/greedy algorithms.

Idealogy

“when life gives you lemons, convexify”

- Sparse use ℓ_1 norm, Low-rank use nuclear norm, atomic norms, etc.

convex algorithms are not perfect

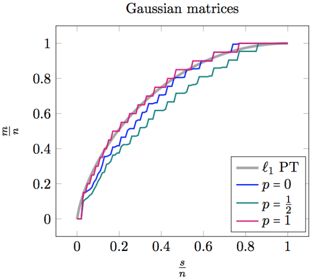
convex algorithms are not perfect

- Computation and memory: convex programs maybe inefficient

convex algorithms are not perfect

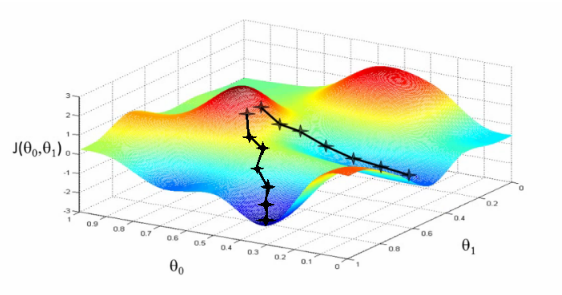
- Computation and memory: convex programs maybe inefficient

- Sometimes convex programs are inefficient in capturing the “structure” (usually require more samples)



Just follow the gradient?

In practice local search heuristics work really well..



Non-convex optimization is difficult!

- Non-convex optimization is quite tricky!

Non-convex optimization is difficult!

- Non-convex optimization is quite tricky!
-

$$f(\boldsymbol{x}) = \sum_{i,j=1}^{n} Q_{ij} x_i^2 x_j^2 \quad \nabla f(0) = 0 \quad \text{for all } Q$$

- Checking if 0 is a local minimum is NP-hard!

Going beyond worse case

Two stories with a common theme.

With randomized coefficients (e.g. functions of Gaussians) local search heuristics work.

- Story I: Solving quadratic equations (nonconvex objective)
- Story II: Linear inverse problems (nonconvex constraints)

Going beyond worse case

Two stories with a common theme.

With randomized coefficients (e.g. functions of Gaussians) local search heuristics work.

- Story I: Solving quadratic equations (nonconvex objective)
- Story II: Linear inverse problems (nonconvex constraints)

Challenge

Real data is not Gaussian ...

Going beyond worse case

Two stories with a common theme.

With randomized coefficients (e.g. functions of Gaussians) local search heuristics work.

- Story I: Solving quadratic equations (nonconvex objective)
- Story II: Linear inverse problems (nonconvex constraints)

Challenge

Real data is not Gaussian ...

Solution

Generic Chaining

Story I:
Solving Quadratic Equations

Solving quadratic equations

$$y_r = |\langle \boldsymbol{a}_r, \boldsymbol{x} \rangle|^2 \quad r = 1, 2, \ldots, m \quad \Leftrightarrow \quad \boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}|^2$$

Find a feasible point in the intersection of quadratic equations

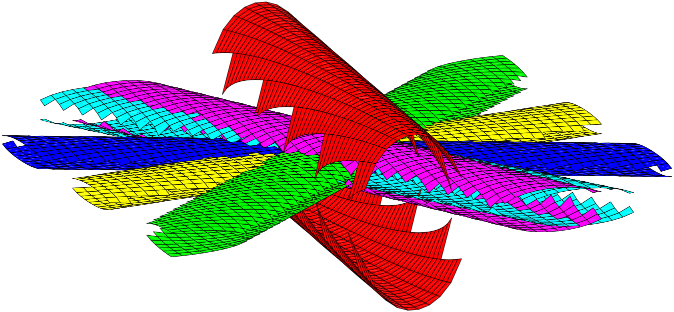
$$\boldsymbol{y}_r = \boldsymbol{x}^* \boldsymbol{A}_r \boldsymbol{x} \quad \text{for} \quad r = 1, 2, \ldots, m.$$

Solving quadratic equations

$$y_r = |\langle \boldsymbol{a}_r, \boldsymbol{x} \rangle|^2 \quad r = 1, 2, \ldots, m \quad \Leftrightarrow \quad \boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}|^2$$

Find a feasible point in the intersection of quadratic equations

$$\boldsymbol{y}_r = \boldsymbol{x}^* \boldsymbol{A}_r \boldsymbol{x} \quad \text{for} \quad r = 1, 2, \ldots, m.$$

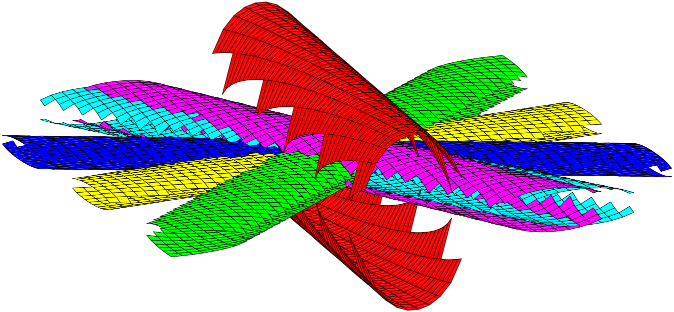


Solving quadratic equations

$$y_r = |\langle \boldsymbol{a}_r, \boldsymbol{x} \rangle|^2 \quad r = 1, 2, \ldots, m \quad \Leftrightarrow \quad \boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}|^2$$

Find a feasible point in the intersection of quadratic equations

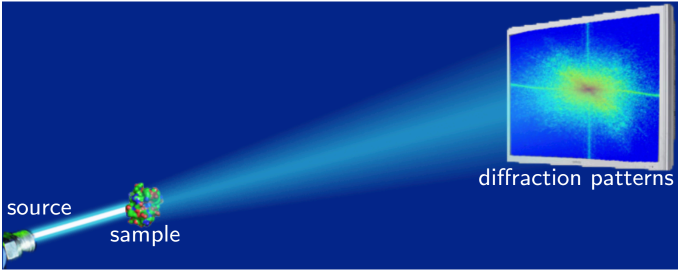
$$\boldsymbol{y}_r = \boldsymbol{x}^* \boldsymbol{A}_r \boldsymbol{x} \quad \text{for} \quad r = 1, 2, \ldots, m.$$



One of the universal forms of combinatorial problems, NP-hard in general.

Missing phase problem

- Detectors only record intensities of diffracted rays (magnitude measurements only!)

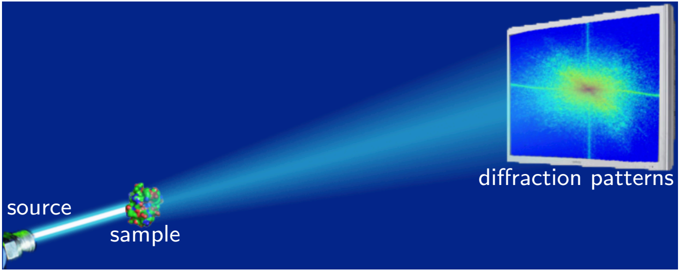


- Fraunhofer diffraction equation $\Rightarrow$ optical field at the detector $\approx$ Fourier transform

$$|\hat{x}(f_1, f_2)|^2 = \left| \int x(t_1, t_2) e^{-2\pi i (f_1 t_1 + f_2 t_2)} dt_1 dt_2 \right|^2$$

Missing phase problem

- Detectors only record intensities of diffracted rays (magnitude measurements only!)



- Fraunhofer diffraction equation $\Rightarrow$ optical field at the detector $\approx$ Fourier transform

$$|\hat{x}(f_1, f_2)|^2 = \left| \int x(t_1, t_2) e^{-2\pi i (f_1 t_1 + f_2 t_2)} dt_1 dt_2 \right|^2$$

Phase Retrieval Problem

How can we recover the phase (or equivalently signal $x(t_1, t_2)$) from $|\hat{x}(f_1, f_2)|$?

Phase retrieval (discrete 1D model)

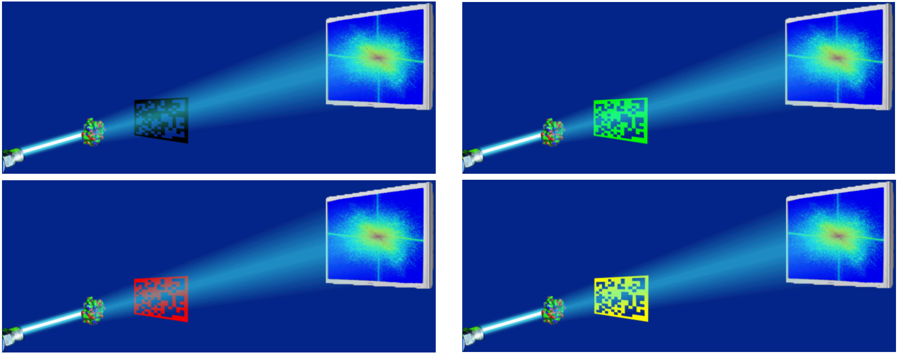
- Phaseless measurements about $\boldsymbol{x} \in \mathbb{C}^n$

$$|\boldsymbol{f}_k^* \boldsymbol{x}|^2 = \boldsymbol{y}_k \quad k \in \{1, 2, \ldots, n\} = [n]$$

  $\boldsymbol{f}_k^*$ is kth row of the DFT matrix.

- Phase retrieval is impossible, inherent ambiguity.

Phase retrieval



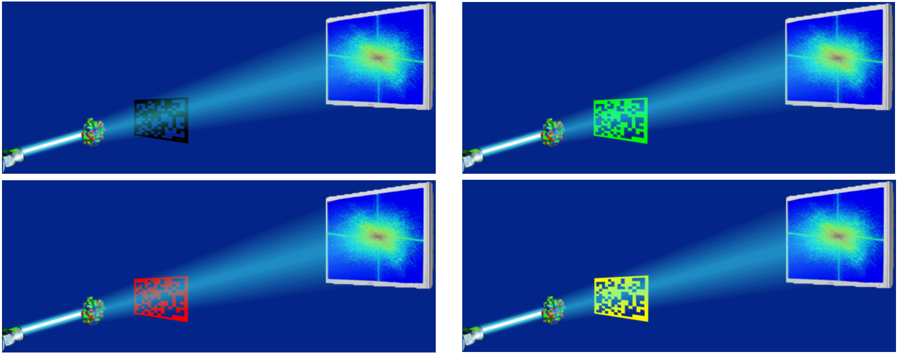
Measurements of the form

$$y_{k\ell} = |\boldsymbol{f}_k^* \boldsymbol{D}_\ell^* \boldsymbol{x}|^2 \quad k = 1, 2, \ldots, n \quad \text{and} \quad \ell = 1, 2, \ldots, L.$$

$$y_r = |\boldsymbol{a}_r^* \boldsymbol{x}|^2 \quad r = 1, 2, \ldots, m.$$

Here $m = nL$ and $\boldsymbol{a}_r = \boldsymbol{D}_\ell \boldsymbol{f}_k, r$ represents (k, ℓ)

Phase retrieval



Measurements of the form

$$y_{k\ell} = |\boldsymbol{f}_k^* \boldsymbol{D}_\ell^* \boldsymbol{x}|^2 \quad k = 1, 2, \ldots, n \quad \text{and} \quad \ell = 1, 2, \ldots, L.$$

$$y_r = |\boldsymbol{a}_r^* \boldsymbol{x}|^2 \quad r = 1, 2, \ldots, m.$$

Here $m = nL$ and $\boldsymbol{a}_r = \boldsymbol{D}_\ell \boldsymbol{f}_k, r$ represents (k, ℓ)

Phase Retrieval by non-convex optimization

Let $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m]$

$$\min_{\boldsymbol{z} \in \mathbb{C}^n} \quad f(\boldsymbol{z}) := \frac{1}{2m} \left\| \boldsymbol{y} - |\boldsymbol{A}^* \boldsymbol{z}|^2 \right\|_{\ell_2}^2$$
$$= \frac{1}{2m} \sum_{r=1}^{m} \left(\boldsymbol{y}_r - |\boldsymbol{a}_r^* \boldsymbol{z}|^2 \right)^2$$

For a vector $\boldsymbol{b} \in \mathbb{C}^n$ by $|b|^2$, I mean Matlab $abs(\boldsymbol{b}.^2)$.

- Pro: operates over vectors much less intensive!
- Con: Non-convex!

Wirtinger Flow (WF)

Algorithm 1 Wirtinger Flow (WF)

Input: Measurements y_r for $r = 1, 2, \ldots, m$.

Initialization (WF-INIT):

Set $\tilde{\boldsymbol{z}}_0$ to be the eigenvector corresponding to the largest eigenvalue of

$$\boldsymbol{Y} = \frac{1}{m} \sum_{r=1}^{m} y_r \boldsymbol{a}_r \boldsymbol{a}_r^*.$$

Set $\boldsymbol{z}_0 = \left(\sqrt{\frac{1}{m} \sum_{r=1}^{m} y_r} \right) \tilde{\boldsymbol{z}}_0$.

Iterations:

for $\tau = 0$ **to** $t - 1$ **do**

Set

$$\boldsymbol{z}_{\tau+1} = \boldsymbol{z}_\tau - \frac{\mu_{\tau+1}}{\|\boldsymbol{z}_0\|_{\ell_2}^2} \left(\frac{1}{m} \sum_{r=1}^{m} \left(|\boldsymbol{a}_r^* \boldsymbol{z}|^2 - y_r \right) (\boldsymbol{a}_r \boldsymbol{a}_r^*) \boldsymbol{z} \right) := \boldsymbol{z}_\tau - \frac{\mu_{\tau+1}}{\|\boldsymbol{z}_0\|_{\ell_2}^2} \nabla f(\boldsymbol{z}_\tau).$$

end for

Output: $\hat{\boldsymbol{x}} = \boldsymbol{z}_t$.

Exact Phase Retrieval by WF (Gaussian Model)

For a vector $\boldsymbol{z} \in \mathbb{C}^n$

$$\mathrm{dist}(\boldsymbol{z}, \boldsymbol{x}) = \min_{\phi \in [0, 2\pi]} \left\| \boldsymbol{z} - e^{i\phi} \boldsymbol{x} \right\|_{\ell_2}.$$

Theorem (Candes, Li, and Soltanolkotabi ('14), Soltanolkotabi ('14))

Assume $m \gtrsim n$. Using $0 \le \mu \le \mu_0 / n$, with high probability

- *Initialization:*

$$\mathit{dist}(\boldsymbol{z}_0, \boldsymbol{x}) \le \sqrt{\frac{5}{6}} \, \|\boldsymbol{x}\|_{\ell_2}.$$

- *After t iterations:*

$$\mathit{dist}(\boldsymbol{z}_t, \boldsymbol{x}) \le e^{-c\mu t} \cdot \mathit{dist}(\boldsymbol{z}_0, \boldsymbol{x}) \le \sqrt{\frac{5}{6}} e^{-c\mu t} \, \|\boldsymbol{x}\|_{\ell_2}.$$

[Chen and Candes 2015] also later established $m \gtrsim n$ via Truncated Wirtinger Flow

Computational complexity

- Initialization: essentially $L \log n$ FFTs$= o(nL \log^2 n)$.
- Iteration update: essentially $2L$ FFTs$= o(nL \log n)$.

Computational complexity

- Initialization: essentially $L \log n$ FFTs$=o(nL \log^2 n)$.
- Iteration update: essentially $2L$ FFTs$=o(nL \log n)$.

Computational Complexity

Near Optimal: Equivalent to a few thousand FFTs.

Computational complexity

- Initialization: essentially $L \log n$ FFTs$=o(nL\log^2 n)$.
- Iteration update: essentially $2L$ FFTs$=o(nL\log n)$.

Computational Complexity

Near Optimal: Equivalent to a few thousand FFTs.

Pros:

- No SVD's.
- No matrix inversions.
- Everything is in $\mathbb{C}^n$ rather than the Lifted space.

A real experiment

Original Image

WF after 500 iterations: Time=221.7697 sec, Relerr=2.5410×10^{-11}

A real experiment

Original Image

WF after 500 iterations: Time=221.7697 sec, Relerr=2.5410×10^{-11}

SDP based methods (PhaseLift,PhaseCut,...) require $\boldsymbol{x}\boldsymbol{x}^*$

A real experiment

Original Image

WF after 500 iterations: Time=221.7697 sec, Relerr=2.5410×10^{-11}

SDP based methods (PhaseLift,PhaseCut,...) require $\boldsymbol{x}\boldsymbol{x}^*$

$(320 \times 1280)^2 \times 8$ Bytes ≈ 0.16777 Tera Bytes

A real experiment

Original Image

WF after 500 iterations: Time=221.7697 sec, Relerr=2.5410×10^{-11}

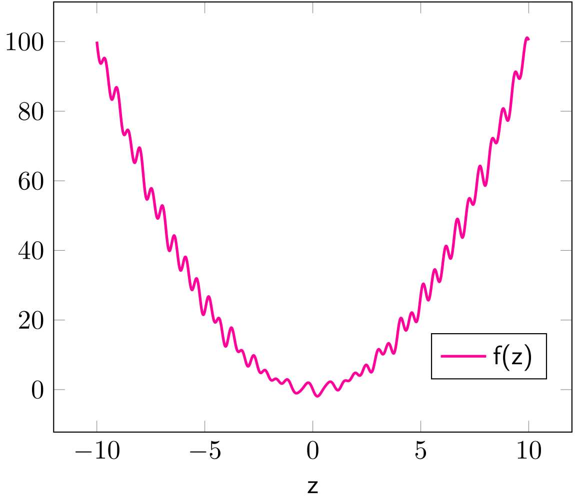
SDP based methods (PhaseLift,PhaseCut,...) require $\boldsymbol{x}\boldsymbol{x}^*$

$(320 \times 1280)^2 \times 8$ Bytes ≈ 0.16777 Tera Bytes

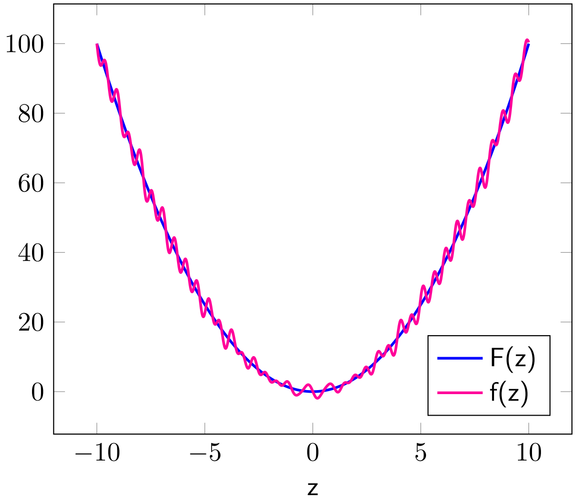
does not even fit into memory!

Proof Sketch

Main Idea: Compare with an easy to analyze function



Main Idea: Compare with an easy to analyze function



For phase retrieval

$$f(\boldsymbol{z}) = \frac{1}{2m} \sum_{r=1}^{m} \left(y_r - |\boldsymbol{a}_r^* \boldsymbol{z}|^2 \right)^2$$

and

$$F(\boldsymbol{z}) := \mathbb{E}[f(\boldsymbol{z})] = \boldsymbol{z}^* \left(\boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^* \right) \boldsymbol{z} + \frac{1}{2} \left(\|\boldsymbol{z}\|_{\ell_2}^2 - \|\boldsymbol{x}\|_{\ell_2}^2 \right)^2$$

For phase retrieval

$$f(\boldsymbol{z}) = \frac{1}{2m}\sum_{r=1}^{m}\left(y_r - |\boldsymbol{a}_r^*\boldsymbol{z}|^2\right)^2$$

and

$$F(\boldsymbol{z}) := \mathbb{E}[f(\boldsymbol{z})] = \boldsymbol{z}^*\left(\boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^*\right)\boldsymbol{z} + \frac{1}{2}\left(\|\boldsymbol{z}\|_{\ell_2}^2 - \|\boldsymbol{x}\|_{\ell_2}^2\right)^2$$

Showing $f(\boldsymbol{z}) \approx F(\boldsymbol{z})$ not sufficient. Need to show this in terms of higher order derivatives.

Regularity Condition

Define

$$P := \{\boldsymbol{x}e^{i\phi} : \phi \in [0, 2\pi]\}.$$

and

$$E(\epsilon) := \{\boldsymbol{z} \in \mathbb{C}^n : \mathsf{dist}(\boldsymbol{z}, P) \leq \epsilon\}.$$

Condition (Regularity Condition)

We say that the function f satisfies the regularity condition or $RC(\alpha, \beta, \epsilon)$ if for all vectors $\boldsymbol{z} \in E(\epsilon)$ we have

$$\mathrm{Re}\left(\left\langle \nabla f(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{x}e^{i\phi(\boldsymbol{z})} \right\rangle\right) \geq \frac{1}{\alpha} \mathit{dist}^2(\boldsymbol{z}, \boldsymbol{x}) + \frac{1}{\beta} \left\| \nabla f(\boldsymbol{z}) \right\|_{\ell_2}^2 .$$

Regularity condition leads to convergence

We will prove that if for all $\boldsymbol{z} \in E(\epsilon)$ we have

$$\mathrm{Re}\left(\left\langle \nabla f(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{x}e^{i\phi(\boldsymbol{z})}\right\rangle\right) \geq \frac{1}{\alpha}\mathsf{dist}^2(\boldsymbol{z}, \boldsymbol{x}) + \frac{1}{\beta}\left\|\nabla f(\boldsymbol{z})\right\|_{\ell_2}^2.$$

then for all $0 < \mu \leq \frac{2}{\beta}$

$$\boldsymbol{z}_+ = \boldsymbol{z} - \mu \nabla f(\boldsymbol{z})$$

obeys

$$\mathsf{dist}^2(\boldsymbol{z}_+, \boldsymbol{x}) \leq \left(1 - \frac{2\mu}{\alpha}\right)\mathsf{dist}^2(\boldsymbol{z}, \boldsymbol{x}).$$

$$\begin{aligned}
\left\|\boldsymbol{z}_+ - \boldsymbol{x}e^{i\phi(\boldsymbol{z}_+)}\right\|_{\ell_2}^2 &\leq \left\|\boldsymbol{z}_+ - \boldsymbol{x}e^{i\phi(\boldsymbol{z})}\right\|_{\ell_2}^2 = \left\|\boldsymbol{z} - \boldsymbol{x}e^{i\phi(\boldsymbol{z})} - \mu\nabla f(\boldsymbol{z})\right\|_{\ell_2}^2 \\
&= \left\|\boldsymbol{z} - \boldsymbol{x}e^{i\phi(\boldsymbol{z})}\right\|_{\ell_2}^2 - 2\mu\,\mathrm{Re}\left(\left\langle \nabla f(\boldsymbol{z}), \left(\boldsymbol{z} - \boldsymbol{x}e^{i\phi(\boldsymbol{z})}\right)\right\rangle\right) + \mu^2\left\|\nabla f(\boldsymbol{z})\right\|_{\ell_2}^2 \\
&\leq \left\|\boldsymbol{z} - \boldsymbol{x}e^{i\phi(\boldsymbol{z})}\right\|_{\ell_2}^2 - 2\mu\left(\frac{1}{\alpha}\left\|\boldsymbol{z} - \boldsymbol{x}e^{i\phi(\boldsymbol{z})}\right\|_{\ell_2}^2 + \frac{1}{\beta}\left\|\nabla f(\boldsymbol{z})\right\|_{\ell_2}^2\right) + \mu^2\left\|\nabla f(\boldsymbol{z})\right\|_{\ell_2}^2 \\
&= \left(1 - \frac{2\mu}{\alpha}\right)\left\|\boldsymbol{z} - \boldsymbol{x}e^{i\phi(\boldsymbol{z})}\right\|_{\ell_2}^2 + \mu\left(\mu - \frac{2}{\beta}\right)\left\|\nabla f(\boldsymbol{z})\right\|_{\ell_2}^2 \\
&\leq \left(1 - \frac{2\mu}{\alpha}\right)\left\|\boldsymbol{z} - \boldsymbol{x}e^{i\phi(\boldsymbol{z})}\right\|_{\ell_2}^2,
\end{aligned}$$

Regularity condition leads to convergence

We will prove that if for all $\boldsymbol{z} \in E(\epsilon)$ we have

$$\mathrm{Re}\left(\left\langle \nabla f(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{x}e^{i\phi(\boldsymbol{z})} \right\rangle\right) \geq \frac{1}{\alpha}\mathsf{dist}^2(\boldsymbol{z}, \boldsymbol{x}) + \frac{1}{\beta}\left\|\nabla f(\boldsymbol{z})\right\|_{\ell_2}^2.$$

then for all $0 < \mu \leq \frac{2}{\beta}$

$$\boldsymbol{z}_+ = \boldsymbol{z} - \mu \nabla f(\boldsymbol{z})$$

obeys

$$\mathsf{dist}^2(\boldsymbol{z}_+, \boldsymbol{x}) \leq \left(1 - \frac{2\mu}{\alpha}\right)\mathsf{dist}^2(\boldsymbol{z}, \boldsymbol{x}).$$

$$\begin{aligned}
\left\|\boldsymbol{z}_+ - \boldsymbol{x}e^{i\phi(\boldsymbol{z}_+)}\right\|_{\ell_2}^2 &\leq \left\|\boldsymbol{z}_+ - \boldsymbol{x}e^{i\phi(\boldsymbol{z})}\right\|_{\ell_2}^2 = \left\|\boldsymbol{z} - \boldsymbol{x}e^{i\phi(\boldsymbol{z})} - \mu\nabla f(\boldsymbol{z})\right\|_{\ell_2}^2 \\
&= \left\|\boldsymbol{z} - \boldsymbol{x}e^{i\phi(\boldsymbol{z})}\right\|_{\ell_2}^2 - 2\mu\,\mathrm{Re}\left(\left\langle \nabla f(\boldsymbol{z}), \left(\boldsymbol{z} - \boldsymbol{x}e^{i\phi(\boldsymbol{z})}\right)\right\rangle\right) + \mu^2\left\|\nabla f(\boldsymbol{z})\right\|_{\ell_2}^2 \\
&\leq \left\|\boldsymbol{z} - \boldsymbol{x}e^{i\phi(\boldsymbol{z})}\right\|_{\ell_2}^2 - 2\mu\left(\frac{1}{\alpha}\left\|\boldsymbol{z} - \boldsymbol{x}e^{i\phi(\boldsymbol{z})}\right\|_{\ell_2}^2 + \frac{1}{\beta}\left\|\nabla f(\boldsymbol{z})\right\|_{\ell_2}^2\right) + \mu^2\left\|\nabla f(\boldsymbol{z})\right\|_{\ell_2}^2 \\
&= \left(1 - \frac{2\mu}{\alpha}\right)\left\|\boldsymbol{z} - \boldsymbol{x}e^{i\phi(\boldsymbol{z})}\right\|_{\ell_2}^2 + \mu\left(\mu - \frac{2}{\beta}\right)\left\|\nabla f(\boldsymbol{z})\right\|_{\ell_2}^2 \\
&\leq \left(1 - \frac{2\mu}{\alpha}\right)\left\|\boldsymbol{z} - \boldsymbol{x}e^{i\phi(\boldsymbol{z})}\right\|_{\ell_2}^2,
\end{aligned}$$

How do we prove the regularity condition?

$$\mathrm{Re}\left(\left\langle \nabla f(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{x}e^{i\phi(\boldsymbol{z})}\right\rangle\right) \geq \frac{1}{\alpha}\mathsf{dist}^2(\boldsymbol{z},\boldsymbol{x}) + \frac{1}{\beta}\left\|\nabla f(\boldsymbol{z})\right\|_{\ell_2}^2 .$$

and

$$\nabla f(\boldsymbol{z}) = \frac{1}{m}\sum_{r=1}^{m}\left(\left|\boldsymbol{a}_r^*\boldsymbol{z}\right|^2 - y_r\right)(\boldsymbol{a}_r^*\boldsymbol{z})\boldsymbol{a}_r$$

Condition (Local Curvature Condition)

$$\mathrm{Re}\left(\left\langle \nabla f(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{x}e^{i\phi(\boldsymbol{z})}\right\rangle\right) \geq \left(\frac{1}{\alpha} + \frac{(1-\delta)}{4}\right)\mathit{dist}^2(\boldsymbol{z},\boldsymbol{x}) + \frac{1}{10m}\sum_{r=1}^{m}\left|\boldsymbol{a}_r^*(\boldsymbol{z} - e^{i\phi(\boldsymbol{z})}\boldsymbol{x})\right|^4$$

Condition (Local Smoothness Condition)

$$\left\|\nabla f(\boldsymbol{z})\right\|_{\ell_2}^2 \leq \beta\left(\frac{(1-\delta)}{4}\mathit{dist}^2(\boldsymbol{z},\boldsymbol{x}) + \frac{1}{10m}\sum_{r=1}^{m}\left|\boldsymbol{a}_r^*(\boldsymbol{z} - e^{i\phi(\boldsymbol{z})}\boldsymbol{x})\right|^4\right).$$

Story II:
Linear inverse problems

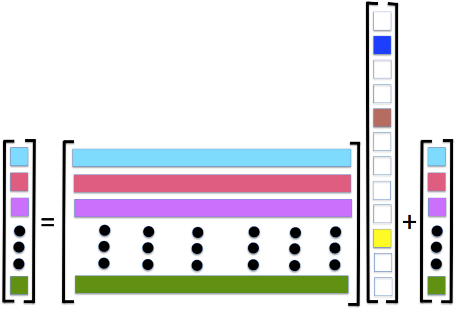
Structured signal recovery/Linear inverse problems

Optimization Problem

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{z}} \quad \frac{1}{2}\sum_{i=1}^{m}\left(y_i - \langle \boldsymbol{a}_i, \boldsymbol{z}\rangle\right)^2 = \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}\right\|_{\ell_2}^2$$

$$\text{subject to} \quad f(\boldsymbol{z}) \le f(\boldsymbol{x})$$

f is a function that enforces structure



$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$$

Minimal data/measurements for exact recovery?

(Review)

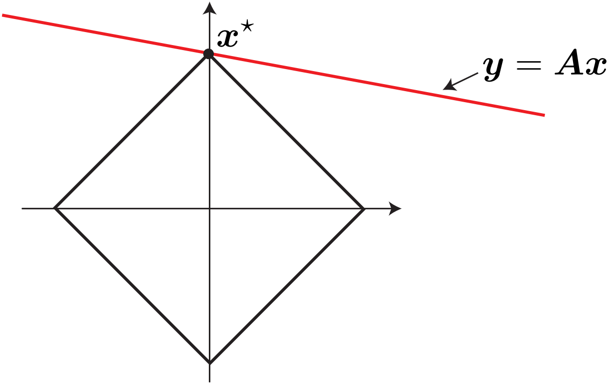
Minimal data/measurements for exact recovery?

$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, $\boldsymbol{y} \in \mathbb{R}^m$, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, and $\boldsymbol{x} \in \mathbb{R}^n$ with $m << n$.

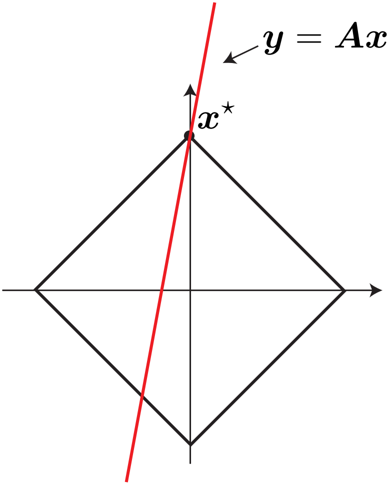
$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{z}}{\text{argmin}} \ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}\|_{\ell_2}^2 \quad \text{subject to} \quad f(\boldsymbol{z}) \leq f(\boldsymbol{x}).$$

When is $\hat{\boldsymbol{x}} = \boldsymbol{x}$? m?

Why does ℓ_1 work?

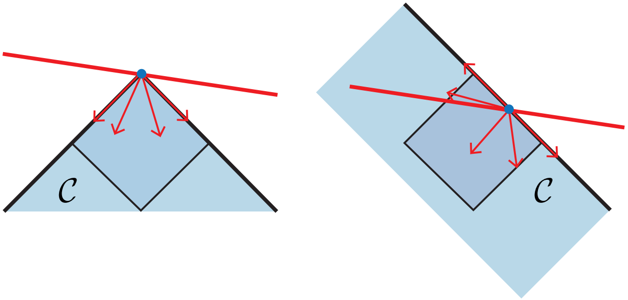


Why ℓ_1 may not always work



Geometry

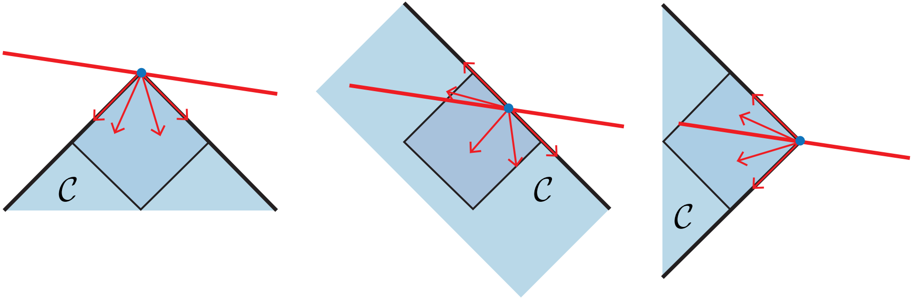
$\mathcal{C} = \{h : \|x + th\| \leq \|x\|$ for some $t > 0\}$ **cone of descent**



Exact recovery if $\mathcal{C} \cap \mathrm{null}(A) = \{0\}$

Geometry

$\mathcal{C} = \{h : \|x + th\| \leq \|x\| \text{ for some } t > 0\}$ **cone of descent**



Exact recovery if $\mathcal{C} \cap \text{null}(A) = \{0\}$

Mean width

$$\omega(\mathcal{T}) = \mathbb{E}_{\boldsymbol{g}}[\sup_{\boldsymbol{x}\in\mathcal{T}} \boldsymbol{g}^T \boldsymbol{x}]$$

- finite set $\mathcal{T}$

Mean width

$$\omega(\mathcal{T}) = \mathbb{E}_{\boldsymbol{g}}[\sup_{\boldsymbol{x}\in\mathcal{T}} \boldsymbol{g}^T\boldsymbol{x}]$$

- finite set $\mathcal{T}$

$$\omega^2(\mathcal{T}) = 2\log|\mathcal{T}|$$

Mean width

$$\omega(\mathcal{T}) = \mathbb{E}_{\boldsymbol{g}}[\sup_{\boldsymbol{x}\in\mathcal{T}} \boldsymbol{g}^T \boldsymbol{x}]$$

- finite set $\mathcal{T}$

$$\omega^2(\mathcal{T}) = 2\log|\mathcal{T}|$$

- $\mathcal{C}$ cone of decent of ℓ_1 norm at an s-sparse signal

Mean width

$$\omega(\mathcal{T}) = \mathbb{E}_{\boldsymbol{g}}[\sup_{\boldsymbol{x}\in\mathcal{T}} \boldsymbol{g}^T\boldsymbol{x}]$$

- finite set $\mathcal{T}$

$$\omega^2(\mathcal{T}) = 2\log|\mathcal{T}|$$

- $\mathcal{C}$ cone of decent of ℓ_1 norm at an s-sparse signal

$$\omega^2(\mathcal{C}\cap\mathbb{S}^{n-1}) \approx 2s\log(n/s)$$

Theorem (Chandrasekaran, Recht, Parrilo, and Willskey 2012)

For i.i.d. normal matrices as long as

$$m \geq m_0(f, \boldsymbol{x}) := \omega^2(\mathcal{C}_f(\boldsymbol{x})),$$

then

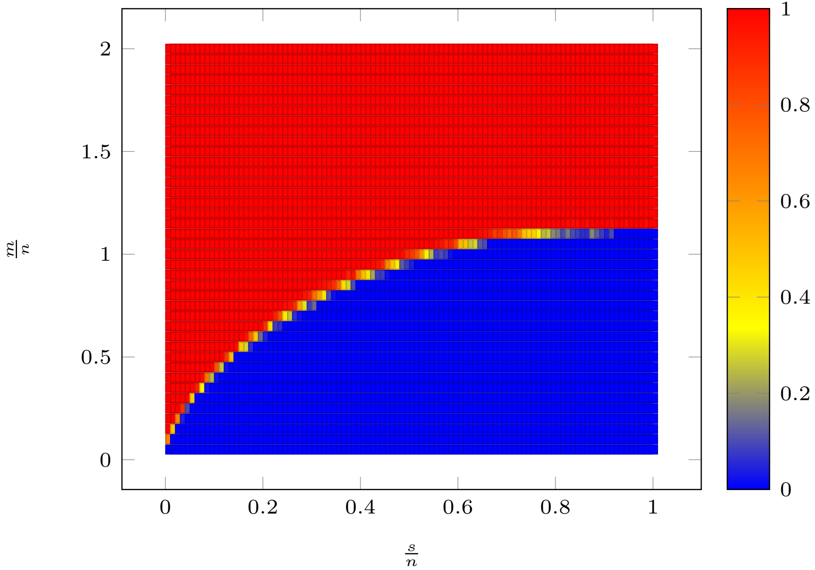
$$\hat{\boldsymbol{x}} = \boldsymbol{x}$$

holds with high probability.

- f is convex and $\boldsymbol{A}$ Gaussian (Chandrasekaran, Recht, Parrilo, Willskey 2012), (Amelunxen, Lotz, McCoy, Tropp 2013), (Stojnic 2009, 2013), Special cases [Donoho-Tanner 2009], many works by (Donoho, Maleki,Montanari 2009) , (Bayati and Montanari 2011)
- sub-Gaussians: [Klartag and Mendelson 2005] , [Mendelson, Pajor and Tomczak-Jaegermann 2007], [Dirksen 2015], [Bayati, Lelarge, Montanari 2014], [Oymak and Tropp 2015]
- non-Gaussians and non-i.i.d. not known for general structures
- Non-Gaussians, non-i.i.d. not known for the most part. Special structures: [Candes, Romberg, Tao 2004],[Rudelson and Vershynin 2007], [Krahmer, Mendelson, Rauhut 2012], [Bourgain, Dirksen, Nelson 2015],

Phase Transitions

Sparse recovery via ℓ_1 minimization

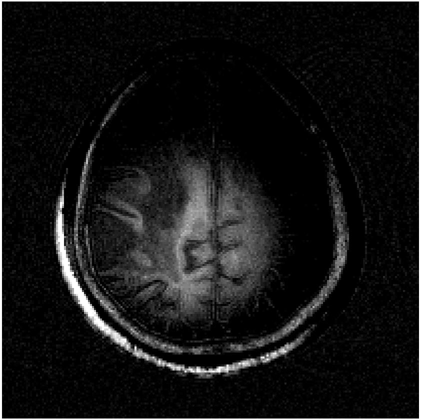


Going beyond convexity

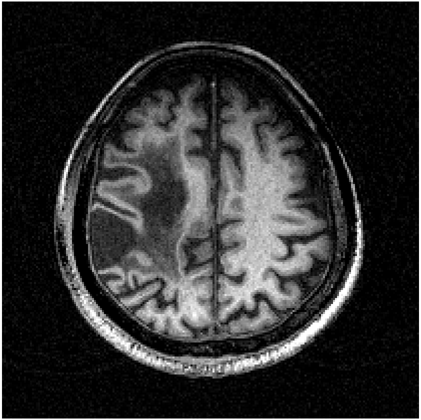
- A huge literature on convex relaxations/greedy methods over the last decade

Going beyond convexity

- A huge literature on convex relaxations/greedy methods over the last decade
- The function that best captures signal structure may not be convex...



real part

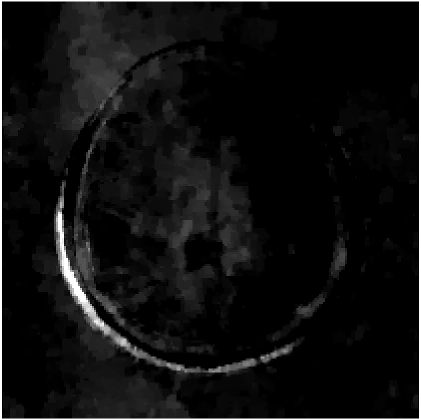


imaginary part

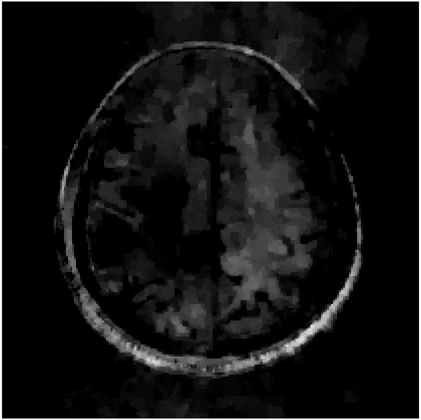
Data courtesy of USC biomedical imaging group

Accelerated MRI via total variation minimization

Undersample by a factor of 3 and use TV reconstruction

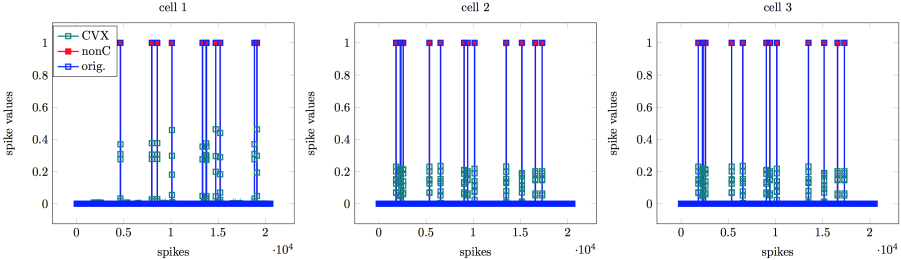


real part



imaginary part

This happens even for sparse signals



Algorithms

Projected Gradient Descent

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{z}}{\text{argmin}} \; \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}\|_{\ell_2}^2 \quad \text{subject to} \quad f(\boldsymbol{z}) \le f(\boldsymbol{x}).$$

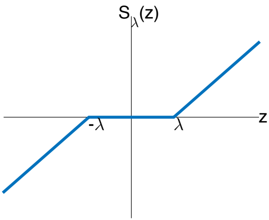
Projection: $\mathcal{P}_{\mathcal{K}}(\boldsymbol{z})$, $\mathcal{K} = \{\boldsymbol{z} \in \mathbb{R}^n : f(\boldsymbol{z}) \le f(\boldsymbol{x})\}$

Projected Gradient Descent

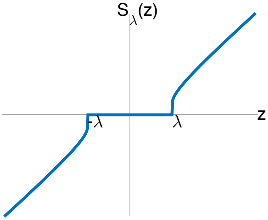
$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{z}}{\text{argmin}}\ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}\|_{\ell_2}^2 \quad \text{subject to} \quad f(\boldsymbol{z}) \leq f(\boldsymbol{x}).$$

Projection: $\mathcal{P}_{\mathcal{K}}(\boldsymbol{z})$, $\mathcal{K} = \{\boldsymbol{z} \in \mathbb{R}^n : f(\boldsymbol{z}) \leq f(\boldsymbol{x})\}$

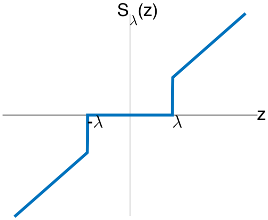
- Start from $\boldsymbol{z}_0 = \boldsymbol{0}$.
- Apply the update $\boldsymbol{z}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{z}_\tau + \mu \boldsymbol{A}^T\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau\right)\right)$



$f(\boldsymbol{z}) = \|\boldsymbol{z}\|_{\ell_1}$



$f(\boldsymbol{z}) = \|\boldsymbol{z}\|_{\ell_{1/2}}^2$



$f(\boldsymbol{z}) = \|\boldsymbol{z}\|_{\ell_0}$

Theory

A simple thought experiment

Reminder: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$.
Algorithm:

- Start with $\boldsymbol{z}_0 = \boldsymbol{0}$,
- Run $\boldsymbol{z}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{z}_\tau + \mu \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)\right)$

A simple thought experiment

Reminder: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$.
Algorithm:

- Start with $\boldsymbol{z}_0 = \boldsymbol{0}$,
- Run $\boldsymbol{z}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{z}_\tau + \mu \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)\right)$

Forget about the projection: $\boldsymbol{z}_{\tau+1} = \boldsymbol{z}_\tau + \mu \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)$

A simple thought experiment

Reminder: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$.
Algorithm:

- Start with $\boldsymbol{z}_0 = \boldsymbol{0}$,
- Run $\boldsymbol{z}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{z}_\tau + \mu \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)\right)$

Forget about the projection: $\boldsymbol{z}_{\tau+1} = \boldsymbol{z}_\tau + \mu \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)$

$$\boldsymbol{z}_{\tau+1} - \boldsymbol{x} = (\boldsymbol{I} - \mu \boldsymbol{A}^* \boldsymbol{A})(\boldsymbol{z}_\tau - \boldsymbol{x}) - \mu \boldsymbol{A}^* \boldsymbol{w}.$$

A simple thought experiment

Reminder: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$.
Algorithm:

- Start with $\boldsymbol{z}_0 = \boldsymbol{0}$,
- Run $\boldsymbol{z}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{z}_\tau + \mu \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)\right)$

Forget about the projection: $\boldsymbol{z}_{\tau+1} = \boldsymbol{z}_\tau + \mu \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)$

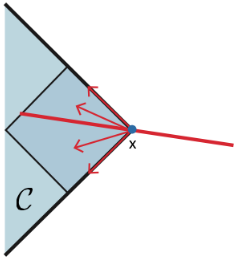
$$\boldsymbol{z}_{\tau+1} - \boldsymbol{x} = (\boldsymbol{I} - \mu \boldsymbol{A}^* \boldsymbol{A})(\boldsymbol{z}_\tau - \boldsymbol{x}) - \mu \boldsymbol{A}^* \boldsymbol{w}.$$

$$\left\| \boldsymbol{z}_{\tau+1} - \boldsymbol{x} \right\|_{\ell_2} \le \left\| \boldsymbol{I} - \mu \boldsymbol{A}^* \boldsymbol{A} \right\| \left\| \boldsymbol{z}_\tau - \boldsymbol{x} \right\|_{\ell_2} + \mu \left\| \boldsymbol{A} \right\| \left\| \boldsymbol{w} \right\|_{\ell_2}.$$

Geometry

Definition (Descent Cone)

$$\mathcal{C}_f(\boldsymbol{x}) = \{\boldsymbol{h} : f(\boldsymbol{x} + \tau\boldsymbol{h}) \leq f(\boldsymbol{x})\}.$$



A deterministic result

Algorithm:

- Start with $\boldsymbol{z}_0 = \mathbf{0}$,
- Run $\boldsymbol{z}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{z}_\tau + \mu \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)\right)$

A deterministic result

Algorithm:

- Start with $\boldsymbol{z}_0 = \boldsymbol{0}$,
- Run $\boldsymbol{z}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{z}_\tau + \mu \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)\right)$

Theorem (Oymak, Soltanolkotabi, and Recht 2016)

For any $\boldsymbol{A}$, and any convex f

$$\|\boldsymbol{z}_\tau - \boldsymbol{x}\|_{\ell_2} \le (\rho(\mu))^\tau \|\boldsymbol{x}\|_{\ell_2} + \frac{1 - (\rho(\mu))^\tau}{1 - \rho(\mu)} \xi_\mu(\boldsymbol{A}) \|\boldsymbol{w}\|_{\ell_2}.$$

- $\rho(\mu)$ is the convergence rate

$$\rho(\mu) := \rho(\mu, \boldsymbol{A}, f, \boldsymbol{x}) = \sup_{\boldsymbol{u},\boldsymbol{v} \in \mathcal{C}_f(\boldsymbol{x}) \cap \mathcal{B}^n} \boldsymbol{u}^* \left(\boldsymbol{I} - \mu \boldsymbol{A}^* \boldsymbol{A}\right) \boldsymbol{v},$$

- $\xi_\mu(\boldsymbol{A})$ is the noise amplification factor

$$\xi_\mu(\boldsymbol{A}) := \xi_\mu(\boldsymbol{A}, f, \boldsymbol{x}, \boldsymbol{w}) = \mu \cdot \sup_{\boldsymbol{v} \in \mathcal{C}_f(\boldsymbol{x}) \cap \mathcal{B}^n} \boldsymbol{v}^* \boldsymbol{A}^* \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_{\ell_2}}.$$

A deterministic result

Algorithm:

- Start with $\boldsymbol{z}_0 = \boldsymbol{0}$,
- Run $\boldsymbol{z}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{z}_\tau + \mu \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)\right)$

Theorem (Oymak, Soltanolkotabi, and Recht 2016)

For any $\boldsymbol{A}$, and any convex f

$$\left\|\boldsymbol{z}_\tau - \boldsymbol{x}\right\|_{\ell_2} \leq \left(\rho(\mu)\right)^\tau \left\|\boldsymbol{x}\right\|_{\ell_2} + \frac{1 - \left(\rho(\mu)\right)^\tau}{1 - \rho(\mu)} \xi_\mu(\boldsymbol{A}) \left\|\boldsymbol{w}\right\|_{\ell_2}.$$

- $\rho(\mu)$ is the convergence rate

$$\rho(\mu) := \rho(\mu, \boldsymbol{A}, f, \boldsymbol{x}) = \sup_{\boldsymbol{u},\boldsymbol{v} \in \mathcal{C}_f(\boldsymbol{x}) \cap \mathcal{B}^n} \boldsymbol{u}^* \left(\boldsymbol{I} - \mu \boldsymbol{A}^* \boldsymbol{A}\right) \boldsymbol{v},$$

- $\xi_\mu(\boldsymbol{A})$ is the noise amplification factor

$$\xi_\mu(\boldsymbol{A}) := \xi_\mu(\boldsymbol{A}, f, \boldsymbol{x}, \boldsymbol{w}) = \mu \cdot \sup_{\boldsymbol{v} \in \mathcal{C}_f(\boldsymbol{x}) \cap \mathcal{B}^n} \boldsymbol{v}^* \boldsymbol{A}^* \frac{\boldsymbol{w}}{\left\|\boldsymbol{w}\right\|_{\ell_2}}.$$

- for non-convex f just replace ρ with 2ρ!

Geometric Convergence!

- For convex f traditional literature will tell you $1/\tau$ convergence!

Geometric Convergence!

- For convex f traditional literature will tell you $1/\tau$ convergence!
- For convex f we have geometric convergence if $\rho < 1$ even though the objective is not strongly convex!

Geometric Convergence!

- For convex f traditional literature will tell you $1/\tau$ convergence!
- For convex f we have geometric convergence if $\rho < 1$ even though the objective is not strongly convex!
- for non-convex f we have geometric convergence to the global optimum when $\rho < \frac{1}{2}$.

Gaussian measurements with different step size

Natural tradeoffs via step sizes:

- greedy: $\mu = 1/m$
- conservative: $\mu \approx \frac{1}{m+n}$

Gaussian measurements with greedy step size

$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$ Algorithm:

- Start with $\boldsymbol{z}_0 = \boldsymbol{0}$,
- Run $\boldsymbol{z}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{z}_\tau + \mu \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)\right)$

Gaussian measurements with greedy step size

$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$ Algorithm:

- Start with $\boldsymbol{z}_0 = \boldsymbol{0}$,
- Run $\boldsymbol{z}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{z}_\tau + \mu \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)\right)$

Theorem (Oymak, Soltanolkotabi, and Recht 2016)

For Gaussian $\boldsymbol{A}$, $\mu = 1/m$, with high probability

$$\|\boldsymbol{z}_\tau - \boldsymbol{x}\|_{\ell_2} \leq \left(\sqrt{8\kappa_f^2 \frac{m_0}{m}}\right)^\tau \|\boldsymbol{z}_0 - \boldsymbol{x}\|_{\ell_2} + \kappa_f \sqrt{\frac{\pi}{2}} \frac{\sqrt{m_0}}{m} \|\boldsymbol{w}\|_{\ell_2}$$

- Sample complexity $m \geq 8{\kappa_f}^2 m_0$
- Linear rate
- $\kappa_f = 1$ for convex f and $\kappa_f = 2$ for non-convex f.
- Works for any function (including non-convex)!!

Gaussian measurements with conservative step size

$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$ Algorithm:

- Start with $\boldsymbol{z}_0 = \boldsymbol{0}$,
- Run $\boldsymbol{z}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{z}_\tau + \mu \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)\right)$

Gaussian measurements with conservative step size

$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$ Algorithm:

- Start with $\boldsymbol{z}_0 = \mathbf{0}$,
- Run $\boldsymbol{z}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{z}_\tau + \mu \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)\right)$

Theorem (2016)

For Gaussian $\boldsymbol{A}$ and convex f, $\mu = \frac{0.99}{(\sqrt{m}+\sqrt{n})^2}$, with high probability

$$\|\boldsymbol{z}_\tau - \boldsymbol{x}\|_{\ell_2} \le \left(1 - \frac{0.3}{m+n}\left(\sqrt{m} - \sqrt{m_0}\right)^2\right)^\tau \cdot \|\boldsymbol{x}\|_{\ell_2} + \frac{3.5}{(1 - \sqrt{\frac{m_0}{m}})^2}\frac{\sqrt{m_0}}{m}\|\boldsymbol{w}\|_{\ell_2}.$$

- Sample complexity $m \ge m_0$
- The rate is now geometric instead of linear

Some historical notes

- for convex and decomposable norms, up to constants (Agarwal, Negahban and Wainwright 2012), convergence upto "statistical accuracy" (Loh and Wainwright 2014)
- approximate message passing algorithm asymptotically achieves $\sqrt{\frac{m_0}{m}}$ for separable and pseudo Lipschitz f (Bayati and Montanari 11)

Nonlinear Observations

Nonlinear observe from a structured signal

$$\boldsymbol{y} = g(\boldsymbol{A}\boldsymbol{x}).$$

How can we recover $\boldsymbol{x}$ from these measurements?

Nonlinear Observations

Nonlinear observe from a structured signal

$$\boldsymbol{y} = g(\boldsymbol{A}\boldsymbol{x}).$$

How can we recover $\boldsymbol{x}$ from these measurements?

Pretend the model was linear and run the same algorithm

- Start with $\boldsymbol{z}_0 = \boldsymbol{0}$,
- Run $\boldsymbol{z}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{z}_\tau + \mu \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)\right)$

Nonlinear Observations

Nonlinear observe from a structured signal

$$\boldsymbol{y} = g(\boldsymbol{A}\boldsymbol{x}).$$

How can we recover $\boldsymbol{x}$ from these measurements?
Pretend the model was linear and run the same algorithm

- Start with $\boldsymbol{z}_0 = \boldsymbol{0}$,
- Run $\boldsymbol{z}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{z}_\tau + \mu \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)\right)$

Theorem (Oymak and Soltanolkotabi 2016)

For Gaussian $\boldsymbol{A}$ and $\mu = 1/m$, with high probability

$$\|\boldsymbol{z}_\tau - \mu\boldsymbol{x}\|_{\ell_2} \le \left(\sqrt{8\kappa_f^2 \frac{m_0}{m}}\right)^\tau \|\boldsymbol{z}_0 - \mu\boldsymbol{x}\|_{\ell_2} + 2\sigma\sqrt{\frac{m_0}{m}},$$

where $\mu := \mathbb{E}[wg(w)]$ and $\sigma^2 := \mathbb{E}[(g(w) - \mu w)^2]$ with $\boldsymbol{w} \sim \mathcal{N}(0,1)$.

Interpretation

$$g(\boldsymbol{A}\boldsymbol{x}) \approx \mu\boldsymbol{A}\boldsymbol{x} + \sigma\boldsymbol{w}.$$

See [Plan and Vershynin 2014] for related results using convex programming.

Non-Gaussians

Subsampled Orthogonal with Random Sign (SORS)

Definition (Subsampled Orthogonal with Random Sign (SORS) matrices)

$\boldsymbol{F} \in \mathbb{R}^{n \times n}$ orthonormal matrix

$$\boldsymbol{F}^* \boldsymbol{F} = \boldsymbol{I} \quad \text{and} \quad \max_{i,j} |\boldsymbol{F}_{ij}| \leq \frac{\Delta}{\sqrt{n}}.$$

subsampled matrix $\boldsymbol{H} \in \mathbb{R}^{m \times n}$ with i.i.d. rows uniformly at random from rows of $\boldsymbol{F}$. $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{D}$, with $\boldsymbol{D} \in \mathbb{R}^{n \times n}$ diagonal sign pattern.

SORS results

$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$ Algorithm:

- Start with $\boldsymbol{z}_0 = \mathbf{0}$,
- Run $\boldsymbol{z}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{z}_\tau + \mu \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)\right)$

SORS results

$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$ Algorithm:

- Start with $\boldsymbol{z}_0 = \mathbf{0}$,
- Run $\boldsymbol{z}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{z}_\tau + \mu \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)\right)$

Theorem (Oymak, Recht and Soltanolkotabi 2016)

$\boldsymbol{A}$ is a SORS matrix

$$m > c_\Delta \cdot (\log n)^4 \cdot m_0,$$

with high probability

$$\|\boldsymbol{z}_\tau - \boldsymbol{x}\|_{\ell_2} \le \left(c_\Delta \frac{m_0}{m} \log^4 n\right)^{\frac{\tau}{2}} \|\boldsymbol{x}\|_{\ell_2}.$$

SORS results

$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$ Algorithm:

- Start with $\boldsymbol{z}_0 = \mathbf{0}$,
- Run $\boldsymbol{z}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{z}_\tau + \mu \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)\right)$

Theorem (Oymak, Recht and Soltanolkotabi 2016)

$\boldsymbol{A}$ *is a SORS matrix*

$$m > c_\Delta \cdot (\log n)^4 \cdot m_0,$$

with high probability

$$\|\boldsymbol{z}_\tau - \boldsymbol{x}\|_{\ell_2} \leq \left(c_\Delta \frac{m_0}{m} \log^4 n\right)^{\frac{\tau}{2}} \|\boldsymbol{x}\|_{\ell_2}.$$

- Optimal (up to logs and constant) for "fast" matrices

SORS results

$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$ Algorithm:

- Start with $\boldsymbol{z}_0 = \mathbf{0}$,
- Run $\boldsymbol{z}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{z}_\tau + \mu \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)\right)$

Theorem (Oymak, Recht and Soltanolkotabi 2016)

$\boldsymbol{A}$ is a SORS matrix

$$m > c_\Delta \cdot (\log n)^4 \cdot m_0,$$

with high probability

$$\|\boldsymbol{z}_\tau - \boldsymbol{x}\|_{\ell_2} \leq \left(c_\Delta \frac{m_0}{m} \log^4 n\right)^{\frac{\tau}{2}} \|\boldsymbol{x}\|_{\ell_2}.$$

- Optimal (up to logs and constant) for "fast" matrices
- For sparse matrices see [Bourgain, Dirksen, Nelson 2014]

SORS results

$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$ Algorithm:

- Start with $\boldsymbol{z}_0 = \boldsymbol{0}$,
- Run $\boldsymbol{z}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{z}_\tau + \mu \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)\right)$

Theorem (Oymak, Recht and Soltanolkotabi 2016)

$\boldsymbol{A}$ is a SORS matrix

$$m > c_\Delta \cdot (\log n)^4 \cdot m_0,$$

with high probability

$$\|\boldsymbol{z}_\tau - \boldsymbol{x}\|_{\ell_2} \leq \left(c_\Delta \frac{m_0}{m} \log^4 n\right)^{\frac{\tau}{2}} \|\boldsymbol{x}\|_{\ell_2}.$$

- Optimal (up to logs and constant) for "fast" matrices
- For sparse matrices see [Bourgain, Dirksen, Nelson 2014]
- Can we get down to $c_\Delta (\log n)(\log m_0)^2$? Maybe by [Haviv and Regev 15]?

Proof sketch

Gordon's lemma

- $\mathcal{C} \in \mathbb{R}^n$
- $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, i.i.d. entries distributed as $\mathcal{N}(0,1)$
- $b_m = \mathbb{E}[\|\boldsymbol{g}\|_{\ell_2}]$ with $\boldsymbol{g} \in \mathbb{R}^m$ distributed as $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_m)$

Gordon's lemma

- $\mathcal{C} \in \mathbb{R}^n$
- $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, i.i.d. entries distributed as $\mathcal{N}(0,1)$
- $b_m = \mathbb{E}[\|\boldsymbol{g}\|_{\ell_2}]$ with $\boldsymbol{g} \in \mathbb{R}^m$ distributed as $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$

Then for all $\boldsymbol{u} \in \mathcal{C}$

- with prob. at least $1 - e^{-\frac{\eta^2}{2}}$

$$\frac{\|\boldsymbol{A}\boldsymbol{u}\|_{\ell_2}}{\|\boldsymbol{u}\|_{\ell_2}} \geq b_m - \big(\omega(\mathcal{C} \cap \mathbb{S}^{n-1}) + \eta\big)$$

- with prob. at least $1 - e^{-\frac{\eta^2}{2}}$

$$\frac{\|\boldsymbol{A}\boldsymbol{u}\|_{\ell_2}}{\|\boldsymbol{u}\|_{\ell_2}} \leq b_m + \big(\omega(\mathcal{C} \cap \mathbb{S}^{n-1}) + \eta\big)$$

Gordon's lemma

- $\mathcal{C} \in \mathbb{R}^n$
- $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, i.i.d. entries distributed as $\mathcal{N}(0,1)$
- $b_m = \mathbb{E}[\|\boldsymbol{g}\|_{\ell_2}]$ with $\boldsymbol{g} \in \mathbb{R}^m$ distributed as $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$

Then for all $\boldsymbol{u} \in \mathcal{C}$

- with prob. at least $1 - e^{-\frac{\eta^2}{2}}$

$$\frac{\|\boldsymbol{A}\boldsymbol{u}\|_{\ell_2}}{\|\boldsymbol{u}\|_{\ell_2}} \geq b_m - \left(\omega(\mathcal{C} \cap \mathbb{S}^{n-1}) + \eta\right)$$

- with prob. at least $1 - e^{-\frac{\eta^2}{2}}$

$$\frac{\|\boldsymbol{A}\boldsymbol{u}\|_{\ell_2}}{\|\boldsymbol{u}\|_{\ell_2}} \leq b_m + \left(\omega(\mathcal{C} \cap \mathbb{S}^{n-1}) + \eta\right)$$

Thus for all $\boldsymbol{u} \in \mathcal{C}$

$$-2\frac{\omega+\eta}{b_m}\|\boldsymbol{u}\|_{\ell_2}^2 + \left(\frac{\omega+\eta}{b_m}\right)^2\|\boldsymbol{u}\|_{\ell_2}^2 \leq \frac{1}{b_m^2}\|\boldsymbol{A}\boldsymbol{u}\|_{\ell_2}^2 - \|\boldsymbol{u}\|_{\ell_2}^2 \leq 2\frac{\omega+\eta}{b_m}\|\boldsymbol{u}\|_{\ell_2}^2 + \left(\frac{\omega+\eta}{b_m}\right)^2$$

Bounding the convergence rate in the convex case

$$\rho(\frac{1}{m}) = \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathcal{C}_f(\boldsymbol{x})\cap\mathbb{S}^{n-1}} \boldsymbol{u}^T \left(\boldsymbol{I} - \frac{1}{m}\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{v}$$

Bounding the convergence rate in the convex case

$$\rho(\frac{1}{m}) = \sup_{\boldsymbol{u},\boldsymbol{v} \in \mathcal{C}_f(\boldsymbol{x}) \cap \mathbb{S}^{n-1}} \boldsymbol{u}^T \left(\boldsymbol{I} - \frac{1}{m} \boldsymbol{A}^T \boldsymbol{A} \right) \boldsymbol{v}$$

Simple algebra

$$\begin{aligned} \boldsymbol{u}^T \left(\boldsymbol{I} - \frac{1}{m} \boldsymbol{A}^T \boldsymbol{A} \right) \boldsymbol{v} =& 2 \left(\left\| \frac{\boldsymbol{u}+\boldsymbol{v}}{2} \right\|_{\ell_2}^2 - \frac{1}{m} \left\| \boldsymbol{A} \left(\frac{\boldsymbol{u}+\boldsymbol{v}}{2} \right) \right\|_{\ell_2}^2 \right) \\ &+ \frac{1}{2} \left(\frac{\|\boldsymbol{A}\boldsymbol{u}\|_{\ell_2}^2}{m} - \|\boldsymbol{u}\|_{\ell_2}^2 \right) \\ &+ \frac{1}{2} \left(\frac{\|\boldsymbol{A}\boldsymbol{v}\|_{\ell_2}^2}{m} - \|\boldsymbol{v}\|_{\ell_2}^2 \right). \end{aligned}$$

Gordon for SORS matrices?

Definition (Subsampled Orthogonal with Random Sign (SORS) matrices)

$\boldsymbol{F} \in \mathbb{R}^{n \times n}$ orthonormal matrix

$$\boldsymbol{F}^* \boldsymbol{F} = \boldsymbol{I} \quad \text{and} \quad \max_{i,j} |\boldsymbol{F}_{ij}| \leq \frac{\Delta}{\sqrt{n}}.$$

subsampled matrix $\boldsymbol{H} \in \mathbb{R}^{m \times n}$ with i.i.d. rows uniformly at random from rows of $\boldsymbol{F}$. $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{D}$, with $\boldsymbol{D} \in \mathbb{R}^{n \times n}$ diagonal sign pattern.

Gordon Holds for SORS matrices

Gordon for SORS matrices?

Definition (Subsampled Orthogonal with Random Sign (SORS) matrices)

$\boldsymbol{F} \in \mathbb{R}^{n \times n}$ orthonormal matrix

$$\boldsymbol{F}^* \boldsymbol{F} = \boldsymbol{I} \quad \text{and} \quad \max_{i,j} |\boldsymbol{F}_{ij}| \leq \frac{\Delta}{\sqrt{n}}.$$

subsampled matrix $\boldsymbol{H} \in \mathbb{R}^{m \times n}$ with i.i.d. rows uniformly at random from rows of $\boldsymbol{F}$. $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{D}$, with $\boldsymbol{D} \in \mathbb{R}^{n \times n}$ diagonal sign pattern.

Gordon Holds for SORS matrices

Theorem (Oymak, Recht, and Soltanolkotabi 2016)

Assume $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is a SORS matrix. Then, with prob. at least $1 - 2e^{-\eta}$

$$\sup_{\boldsymbol{x} \in \mathcal{T}} \left| \|\boldsymbol{A}\boldsymbol{x}\|_{\ell_2}^2 - \|\boldsymbol{x}\|_{\ell_2}^2 \right| \leq \max(\delta, \delta^2) \cdot (\mathit{rad}(\mathcal{T}))^2 \,,$$

as long as $m \geq C\Delta^2(1+\eta)^2(\log n)^4 \frac{\max\left(1, \frac{\omega^2(\mathcal{T})}{(\mathit{rad}(\mathcal{T}))^2}\right)}{\delta^2}$

Open problem

- $\mathcal{C} \in \mathbb{R}^n$
- $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, sub-sampled Fourier multiplied random sign.
- $b_m = \mathbb{E}[\|\boldsymbol{g}\|_{\ell_2}]$ with $\boldsymbol{g} \in \mathbb{R}^m$ distributed as $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_m)$

Open problem

- $\mathcal{C} \in \mathbb{R}^n$
- $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, sub-sampled Fourier multiplied random sign.
- $b_m = \mathbb{E}[\|\boldsymbol{g}\|_{\ell_2}]$ with $\boldsymbol{g} \in \mathbb{R}^m$ distributed as $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$

Conjecture

- *with prob. at least* $1 - ???$

$$\inf_{\boldsymbol{u} \in \mathcal{C}} \frac{\|\boldsymbol{A}\boldsymbol{u}\|_{\ell_2}}{\|\boldsymbol{u}\|_{\ell_2}} \geq b_m - \left(\omega(\mathcal{C} \cap \mathbb{S}^{n-1}) + \eta\right)$$

- *with prob. at least* $1 - ????$

$$\sup_{\boldsymbol{u} \in \mathcal{C}} \frac{\|\boldsymbol{A}\boldsymbol{u}\|_{\ell_2}}{\|\boldsymbol{u}\|_{\ell_2}} \leq \left(b_m + \left(\omega(\mathcal{C} \cap \mathbb{S}^{n-1}) + \eta\right)\right) \sqrt{\log n} ????$$

This would settle universality conjecture in compressive sensing!

SORS matrices: from RIP to JL

- Restricted Isometry Property-RIP(s,δ)

$$\sup_{\boldsymbol{x}:\|\boldsymbol{x}\|_{\ell_0}\le s} \left| \|\boldsymbol{F}\boldsymbol{x}\|_{\ell_2}^2 - \|\boldsymbol{x}\|_{\ell_2}^2 \right| \le \max(\delta, \delta^2)\, \|\boldsymbol{x}\|_{\ell_2}^2$$

Theorem (Discrete JL embedding via RIP, Krahmer-Ward 2012)

Assume $\mathcal{T} \in \mathbb{R}^n$ finite points. Suppose

- $\boldsymbol{H} \in \mathbb{R}^{m\times n}$ *obeys RIP(s,δ)*
- $s \lesssim \log(|\mathcal{T}|)$ *and* $0 < \delta \le \frac{\epsilon}{4}$

Then $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{D}$ with $\boldsymbol{D}$ diagonal sign pattern obeys

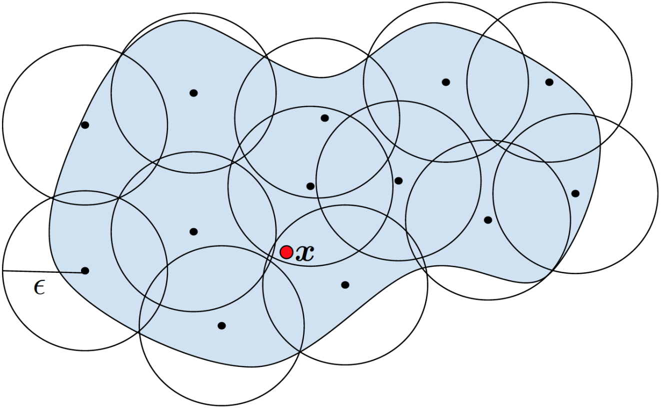
$$\sup_{\boldsymbol{x}\in\mathcal{T}} \left| \|\boldsymbol{A}\boldsymbol{x}\|_{\ell_2}^2 - \|\boldsymbol{x}\|_{\ell_2}^2 \right| \le \max(\epsilon, \epsilon^2)\, \|\boldsymbol{x}\|_{\ell_2}^2 \,,$$

with high probability.

From JL to Gordon?

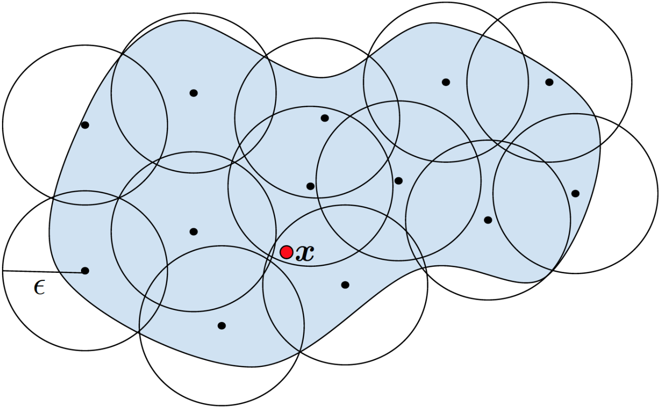
From JL to Gordon?

First attempt: covering



From JL to Gordon?

First attempt: covering

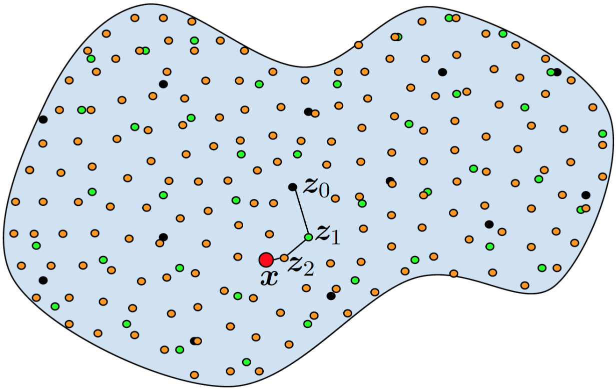


$$m \gtrsim \frac{\log |\mathcal{N}|}{\delta^2} \log^4 n \sim \frac{n}{m} \frac{\omega^2(\mathcal{T})}{\delta^4} \log^4 n \quad \Leftrightarrow \quad m \gtrsim \sqrt{n} \frac{\omega(\mathcal{T})}{\delta^2} \log^2 n,$$

From JL to Gordon?

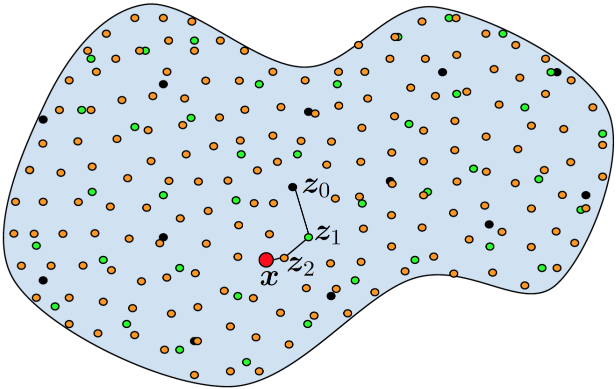
From JL to Gordon?

Second attempt: generic chaining



From JL to Gordon?

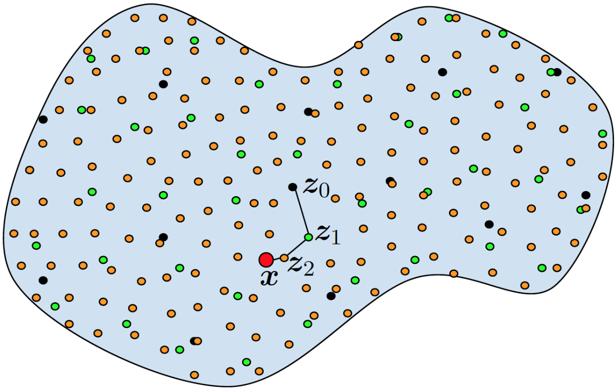
Second attempt: generic chaining



- Successive approximations of size $|\mathcal{T}_\ell| = 2^{2^\ell}$

From JL to Gordon?

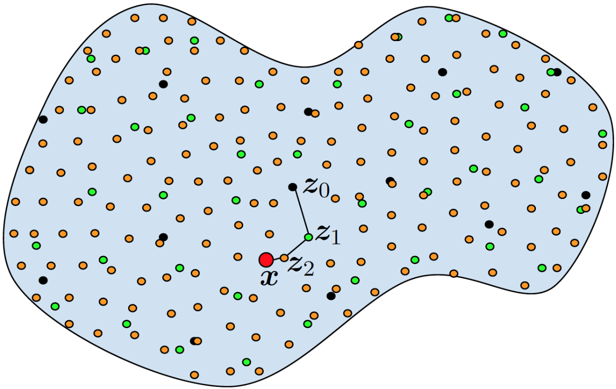
Second attempt: generic chaining



- Successive approximations of size $|\mathcal{T}_\ell| = 2^{2^\ell}$
- Different distortion levels $\delta_\ell = 2^{\ell/2} \frac{\delta}{\omega(\mathcal{T})}$

From JL to Gordon?

Second attempt: generic chaining



- Successive approximations of size $|\mathcal{T}_\ell| = 2^{2^\ell}$
- Different distortion levels $\delta_\ell = 2^{\ell/2} \frac{\delta}{\omega(\mathcal{T})}$

$$m \gtrsim \max_{\ell=1,2,\ldots,L} \frac{\log|\mathcal{T}_\ell|}{\delta_\ell^2} \log^4 n = \max_{\ell=1,2,\ldots,L} \frac{\log|2^{2^\ell}|}{2^\ell \frac{\delta^2}{\omega^2(\mathcal{T})}} \log^4 n = \frac{\omega^2(\mathcal{T})}{\delta^2} \sim \log^4 n \frac{\omega^2(\mathcal{T})}{\delta^2},$$

Is the over all distortion small enough?

After some work Discrete JL implies that with high probability

- For all $\boldsymbol{v} \in \mathcal{T}_{\ell-1} \cup \mathcal{T}_{\ell} \cup (\mathcal{T}_{\ell-1} - \mathcal{T}_{\ell})$,

$$\|\boldsymbol{A}\boldsymbol{v}\|_{\ell_2} \leq (1+\delta_\ell) \|\boldsymbol{v}\|_{\ell_2}.$$

- For all $\boldsymbol{v} \in \mathcal{T}_{\ell-1} \cup \mathcal{T}_{\ell} \cup (\mathcal{T}_{\ell-1} - \mathcal{T}_{\ell})$,

$$\left| \|\boldsymbol{A}\boldsymbol{v}\|_{\ell_2}^2 - \|\boldsymbol{v}\|_{\ell_2}^2 \right| \leq \max\left(\delta_\ell, \delta_\ell^2\right) \cdot \|\boldsymbol{v}\|_{\ell_2}^2.$$

- For all $\boldsymbol{u} \in \mathcal{T}_{\ell-1}$ and $\boldsymbol{v} \in \mathcal{T}_{\ell} - \{\boldsymbol{u}\} := \{\boldsymbol{y} - \boldsymbol{u} : \boldsymbol{y} \in \mathcal{T}_{\ell}\}$,

$$|\boldsymbol{u}^*\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{v} - \boldsymbol{u}^*\boldsymbol{v}| \leq \max\left(\delta_\ell, \delta_\ell^2\right) \cdot \|\boldsymbol{u}\|_{\ell_2} \|\boldsymbol{v}\|_{\ell_2}.$$

where

$$\delta_\ell = 2^{\ell/2} \frac{\delta}{\omega(\mathcal{T})}$$

Is the over all distortion small enough?

We are interested in bounding $|\,\|\boldsymbol{A}\boldsymbol{x}\|_{\ell_2}^2 - \|\boldsymbol{x}\|_{\ell_2}^2\,|$ for all $\boldsymbol{x} \in \mathcal{T}$.
Define $\tilde{L} = \max\left(0, \lfloor 2\log_2\left(\frac{\omega(\mathcal{T})}{\delta}\right)\rfloor\right)$

$$
\begin{aligned}
|\,\|\boldsymbol{A}\boldsymbol{x}\|_{\ell_2}^2 - \|\boldsymbol{x}\|_{\ell_2}^2\,| \leq & \left|\|\boldsymbol{A}\boldsymbol{z}_{\tilde{L}}\|_{\ell_2}^2 - \|\boldsymbol{z}_{\tilde{L}}\|_{\ell_2}^2\right| \\
& + \left|\|\boldsymbol{A}\boldsymbol{x}\|_{\ell_2}^2 - \|\boldsymbol{A}\boldsymbol{z}_{\tilde{L}}\|_{\ell_2}^2\right| + \left|\|\boldsymbol{x}\|_{\ell_2}^2 - \|\boldsymbol{z}_{\tilde{L}}\|_{\ell_2}^2\right| \\
\leq & \sum_{\ell=1}^{\tilde{L}} \left(\left|\|\boldsymbol{A}\boldsymbol{z}_\ell\|_{\ell_2}^2 - \|\boldsymbol{z}_\ell\|_{\ell_2}^2\right| - \left|\|\boldsymbol{A}\boldsymbol{z}_{\ell-1}\|_{\ell_2}^2 - \|\boldsymbol{z}_{\ell-1}\|_{\ell_2}^2\right|\right) \\
& + \left|\|\boldsymbol{A}\boldsymbol{x}\|_{\ell_2}^2 - \|\boldsymbol{A}\boldsymbol{z}_{\tilde{L}}\|_{\ell_2}^2\right| + \left|\|\boldsymbol{x}\|_{\ell_2}^2 - \|\boldsymbol{z}_{\tilde{L}}\|_{\ell_2}^2\right| \\
& + \left|\|\boldsymbol{A}\boldsymbol{z}_0\|_{\ell_2}^2 - \|\boldsymbol{z}_0\|_{\ell_2}^2\right|.
\end{aligned}
$$

Is the over all distortion small enough? (First term)

For the first term

$$\left(\left|\|\boldsymbol{A}\boldsymbol{z}_\ell\|_{\ell_2}^2 - \|\boldsymbol{z}_\ell\|_{\ell_2}^2\right| - \left|\|\boldsymbol{A}\boldsymbol{z}_{\ell-1}\|_{\ell_2}^2 - \|\boldsymbol{z}_{\ell-1}\|_{\ell_2}^2\right|\right) \le 10 e_{\ell-1}\delta_\ell,$$

where $e_\ell =$dist$(\boldsymbol{x}, \mathcal{T}_\ell)$. Then

$$\begin{aligned}\sum_{\ell=1}^{\tilde{L}} \left(\left|\|\boldsymbol{A}\boldsymbol{z}_\ell\|_{\ell_2}^2 - \|\boldsymbol{z}_\ell\|_{\ell_2}^2\right| - \left|\|\boldsymbol{A}\boldsymbol{z}_{\ell-1}\|_{\ell_2}^2 - \|\boldsymbol{z}_{\ell-1}\|_{\ell_2}^2\right|\right) \le & 10\frac{\delta}{\omega(\mathcal{T})}\left(\sum_{\ell=1}^{\tilde{L}} 2^{\ell/2} e_{\ell-1}\right) \\ = & 10\sqrt{2}\frac{\delta}{\omega(\mathcal{T})}\left(\sum_{\ell=0}^{\tilde{L}-1} 2^{\ell/2} e_\ell\right) \\ = & 10\sqrt{2}\frac{\delta}{\omega(\mathcal{T})}\gamma_2(\mathcal{T}) \\ \le & c\delta.\end{aligned}$$

Is the over all distortion small enough? (second term)

$$
\begin{aligned}
\left| \|\boldsymbol{A}\boldsymbol{x}\|_{\ell_2} - \left\|\boldsymbol{A}\boldsymbol{z}_{\tilde{L}}\right\|_{\ell_2} \right| &= \left| \|\boldsymbol{A}\boldsymbol{x}\|_{\ell_2} - \|\boldsymbol{A}\boldsymbol{z}_L\|_{\ell_2} + \|\boldsymbol{A}\boldsymbol{z}_L\|_{\ell_2} - \left\|\boldsymbol{A}\boldsymbol{z}_{\tilde{L}}\right\|_{\ell_2} \right| \\
&\leq \|\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{z}_L)\|_{\ell_2} + \left\|\boldsymbol{A}(\boldsymbol{z}_L - \boldsymbol{z}_{\tilde{L}})\right\|_{\ell_2} \\
&\leq \|\boldsymbol{A}\| \, \|\boldsymbol{x} - \boldsymbol{z}_L\|_{\ell_2} + \left\| \sum_{\ell=\tilde{L}+1}^{L} \boldsymbol{A}(\boldsymbol{z}_\ell - \boldsymbol{z}_{\ell-1}) \right\|_{\ell_2} \\
&\leq \left(\frac{1}{4} 2^{\frac{L}{2}} \frac{\delta}{\omega(\mathcal{T})} + 1 \right) e_L + \sum_{\ell=\tilde{L}+1}^{L} \left\| \boldsymbol{A}(\boldsymbol{z}_\ell - \boldsymbol{z}_{\ell-1}) \right\|_{\ell_2} \\
&\leq \left(\frac{1}{4} 2^{\frac{L}{2}} \frac{\delta}{\omega(\mathcal{T})} + 1 \right) e_L + \sum_{\ell=\tilde{L}+1}^{L} \left(1 + 2^{\ell/2} \frac{\delta}{\omega(\mathcal{T})} \right) \left\| \boldsymbol{z}_\ell - \boldsymbol{z}_{\ell-1} \right\|_{\ell_2} \\
&\leq \frac{5}{4} 2^{L/2} \frac{\delta}{\omega(\mathcal{T})} e_L + \sum_{\ell=\tilde{L}+1}^{L} 2^{\ell/2+1} \frac{\delta}{\omega(\mathcal{T})} \left\| \boldsymbol{z}_\ell - \boldsymbol{z}_{\ell-1} \right\|_{\ell_2} \\
&\leq \frac{5}{4} \frac{\delta}{\omega(\mathcal{T})} 2^{L/2} e_L + 4\sqrt{2} \frac{\delta}{\omega(\mathcal{T})} \sum_{\ell=\tilde{L}+1}^{L} 2^{(\ell-1)/2} e_{\ell-1} \\
&\leq 4\sqrt{2} \frac{\delta}{\omega(\mathcal{T})} \left(\sum_{\ell=\tilde{L}}^{L} 2^{\ell/2} e_\ell \right) \\
&\leq 4\sqrt{2} \frac{\delta}{\omega(\mathcal{T})} \gamma_2(\mathcal{T}).
\end{aligned}
$$

Is the over all distortion small enough? (second term cont.)

$$\begin{aligned}
\left| \|\boldsymbol{A}\boldsymbol{x}\|_{\ell_2}^2 - \left\|\boldsymbol{A}\boldsymbol{z}_{\tilde{L}}^2\right\|_{\ell_2} \right| &\leq \left| \|\boldsymbol{A}\boldsymbol{x}\|_{\ell_2} - \|\boldsymbol{A}\boldsymbol{z}_{\tilde{L}}\|_{\ell_2} \right| \left| \|\boldsymbol{A}\boldsymbol{x}\|_{\ell_2} + \|\boldsymbol{A}\boldsymbol{z}_{\tilde{L}}\|_{\ell_2} \right| \\
&\leq \left| \|\boldsymbol{A}\boldsymbol{x}\|_{\ell_2} - \|\boldsymbol{A}\boldsymbol{z}_{\tilde{L}}\|_{\ell_2} \right|^2 + 2 \left| \|\boldsymbol{A}\boldsymbol{x}\|_{\ell_2} - \|\boldsymbol{A}\boldsymbol{z}_{\tilde{L}}\|_{\ell_2} \right| \|\boldsymbol{A}\boldsymbol{z}_{\tilde{L}}\|_{\ell_2} \\
&\leq 32 \frac{\delta^2}{\omega^2(\mathcal{T})} \gamma_2^2(\mathcal{T}) + 16\sqrt{2} \frac{\delta}{\omega(\mathcal{T})} \gamma_2(\mathcal{T}) \\
&\leq c\delta.
\end{aligned}$$

Is the over all distortion small enough? (third term)

Easy

$$\begin{aligned} \left| \, \|\boldsymbol{A}\boldsymbol{z}_0\|_{\ell_2}^2 - \|\boldsymbol{z}_0\|_{\ell_2}^2 \, \right| &\le \max\left(\frac{\delta}{\omega(\mathcal{T})}, \left(\frac{\delta}{\omega(\mathcal{T})}\right)^2\right) \|\boldsymbol{z}_0\|_{\ell_2}^2 \\ &\le \max\left(\frac{\delta}{\omega(\mathcal{T})}, \left(\frac{\delta}{\omega(\mathcal{T})}\right)^2\right). \end{aligned}$$

Is the failure probability small enough?

After some work Discrete JL implies that with high probability

- For all $\boldsymbol{v} \in \mathcal{T}_{\ell-1} \cup \mathcal{T}_\ell \cup (\mathcal{T}_{\ell-1} - \mathcal{T}_\ell)$,

$$\|\boldsymbol{A}\boldsymbol{v}\|_{\ell_2} \leq (1+\delta_\ell)\,\|\boldsymbol{v}\|_{\ell_2}\,.$$

- For all $\boldsymbol{v} \in \mathcal{T}_{\ell-1} \cup \mathcal{T}_\ell \cup (\mathcal{T}_{\ell-1} - \mathcal{T}_\ell)$,

$$\left|\,\|\boldsymbol{A}\boldsymbol{v}\|_{\ell_2}^2 - \|\boldsymbol{v}\|_{\ell_2}^2\,\right| \leq \max\left(\delta_\ell, \delta_\ell^2\right) \cdot \|\boldsymbol{v}\|_{\ell_2}^2\,.$$

- For all $\boldsymbol{u} \in \mathcal{T}_{\ell-1}$ and $\boldsymbol{v} \in \mathcal{T}_\ell - \{\boldsymbol{u}\} := \{\boldsymbol{y} - \boldsymbol{u} : \boldsymbol{y} \in \mathcal{T}_\ell\}$,

$$|\boldsymbol{u}^*\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{v} - \boldsymbol{u}^*\boldsymbol{v}| \leq \max\left(\delta_\ell, \delta_\ell^2\right) \cdot \|\boldsymbol{u}\|_{\ell_2}\|\boldsymbol{v}\|_{\ell_2}\,.$$

where

$$\delta_\ell = 2^{\ell/2}\frac{\delta}{\omega(\mathcal{T})}$$

The overall probability

$$\sum_{\ell=1}^{L} e^{-\ell(\eta+1)} \leq \sum_{\ell=1}^{\infty} e^{-\ell(\eta+1)} = \frac{e^{-(\eta+1)}}{1-e^{-(\eta+1)}} \leq e^{-\eta},$$

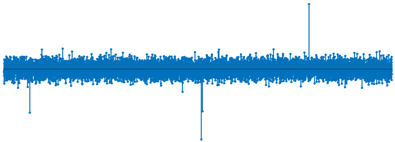
From de-noising to compressed sensing

Minimal number of data?

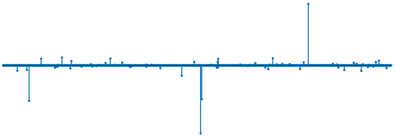
Minimal number of data?

Answer: Intimately related to de-noising capability of the function

Before Thresholding: $\boldsymbol{z}_\tau + \boldsymbol{A}^T\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau\right)$



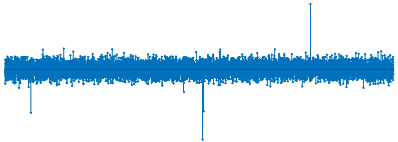
After Thresholding: $\mathcal{P}_{\mathcal{K}}\left(\boldsymbol{z}_\tau + \boldsymbol{A}^T\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau\right)\right)$



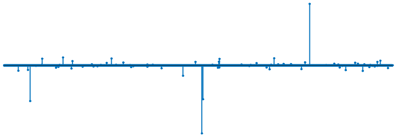
Minimal number of data?

Answer: Intimately related to de-noising capability of the function

Before Thresholding: $\boldsymbol{z}_\tau + \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)$



After Thresholding: $\mathcal{P}_\mathcal{K}\left(\boldsymbol{z}_\tau + \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{z}_\tau)\right)$



Conclusion

Intuitively better de-noiser should work with less data

Add Gaussian noise to parameter and then de-noise!

Add Gaussian noise to parameter and then de-noise!

Theorem (Oymak, Recht and Soltanolkotabi 2016)

For Gaussian $\boldsymbol{A}$, and convex $\mathcal{K}$ minimal number of data is

$$\max_{\sigma} \frac{\mathbb{E}\,\|\mathcal{P}_{\mathcal{K}}(\boldsymbol{x}+\sigma\boldsymbol{z})-\boldsymbol{x}\|_{\ell_2}^2}{\sigma^2} = m_0$$

For non-convex $\mathcal{K}$

$$\max_{\sigma} \frac{\mathbb{E}\,\|\mathcal{P}_{\mathcal{K}}(\boldsymbol{x}+\sigma\boldsymbol{z})-\boldsymbol{x}\|_{\ell_2}^2}{\sigma^2} \leq 4m_0$$

$m_0 \propto \#$ params e.g. $(2s+1)\log(n/s)$

Add Gaussian noise to parameter and then de-noise!

Theorem (Oymak, Recht and Soltanolkotabi 2016)

For Gaussian $\boldsymbol{A}$, and convex $\mathcal{K}$ minimal number of data is

$$\max_{\sigma} \frac{\mathbb{E}\,\|\mathcal{P}_{\mathcal{K}}(\boldsymbol{x}+\sigma\boldsymbol{z})-\boldsymbol{x}\|_{\ell_2}^2}{\sigma^2} = m_0$$

For non-convex $\mathcal{K}$

$$\max_{\sigma} \frac{\mathbb{E}\,\|\mathcal{P}_{\mathcal{K}}(\boldsymbol{x}+\sigma\boldsymbol{z})-\boldsymbol{x}\|_{\ell_2}^2}{\sigma^2} \leq 4m_0$$

$m_0 \propto \#$ params e.g. $(2s+1)\log(n/s)$ We establish a conjecture of Donoho, Johnston and Montanari 2011 up to a small constant.

Real experiment

Related AMP simulations

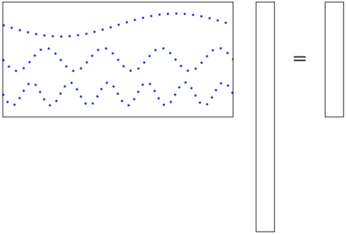
(Montanari IWT'12 plenary)

(Metzler-Maleki-Barinuk '14)

- Signal (1,000,080 pixels)

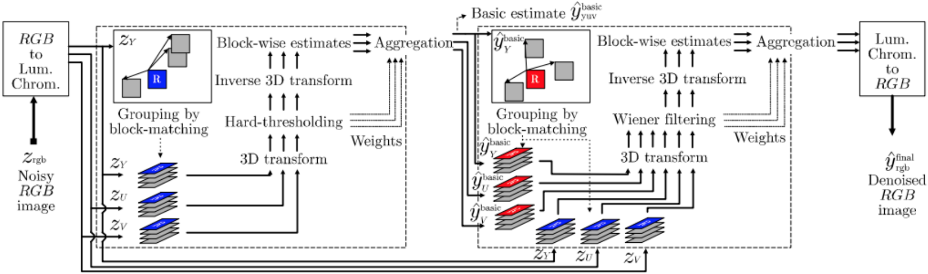
- Signal (1,000,080 pixels)

- Response (324,000 Measurements)



Good image denoiser

CBM3D



Is CBM3D a good denoiser

Signal

Is CBM3D a good denoiser

Signal	Signal+Gaussian noise
	

Is CBM3D a good denoiser

Signal	Signal+Gaussian noise	Result of CBMD3
		

Iteration 1:

Before Thresholding:

Relerr$=0.70687$

After Thresholding:

Relerr$=0.57926$

Original signal $\boldsymbol{x}$:

Iteration 2:

Before Thresholding:

Relerr$=0.54722$

After Thresholding:

Relerr$=0.45153$

Original signal $\boldsymbol{x}$:

Iteration 5:

Before Thresholding:	After Thresholding:	Original signal x:
		
Relerr$=0.19012$	Relerr$=0.15806$	

Iteration 10:

Before Thresholding:

Relerr$=0.062034$

After Thresholding:

Relerr$=0.081282$

Original signal $\boldsymbol{x}$:

Iteration 20:

Before Thresholding:

Relerr$=0.04864$

After Thresholding:

Relerr$=0.061786$

Original signal $\boldsymbol{x}$:

Iteration 40:

Before Thresholding:

Relerr$=0.027613$

After Thresholding:

Relerr$=0.031075$

Original signal $\boldsymbol{x}$:

Iteration 100:

Before Thresholding:

Relerr$=0.015708$

After Thresholding:

Relerr$=0.015732$

Original signal $\boldsymbol{x}$:

Iteration 200:

Before Thresholding:

Relerr$=0.015603$

After Thresholding:

Relerr$=0.015618$

Original signal $\boldsymbol{x}$:

Open Problem

Not very precise conjecture

Conjecture

Assume that for a mapping $\mathcal{S}:\mathbb{R}^m\rightarrow\mathbb{R}^m$ we have

$$\left\|\mathcal{S}(\boldsymbol{u})-\mathcal{S}(\boldsymbol{v})\right\|_{\ell_2}\le L\left\|\boldsymbol{u}-\boldsymbol{v}\right\|_{\ell_2}.$$

show that for all $\boldsymbol{z}\in$Range($\mathcal{S}$)

$$\left\|\mathcal{S}\left(\boldsymbol{x}+\left(\boldsymbol{I}-\frac{1}{m}\boldsymbol{A}^*\boldsymbol{A}\right)(\boldsymbol{z}-\boldsymbol{x})\right)-\boldsymbol{x}\right\|_{\ell_2}\approx\left\|\mathcal{S}\left(\boldsymbol{x}+\frac{1}{\sqrt{m}}\left\|\boldsymbol{z}-\boldsymbol{x}\right\|_{\ell_2}\boldsymbol{g}\right)-\boldsymbol{x}\right\|_{\ell_2}$$

as long as

$$m\ge\max_{\sigma}\frac{\mathbb{E}\left\|\mathcal{S}(\boldsymbol{x}+\sigma\boldsymbol{z})-\boldsymbol{x}\right\|_{\ell_2}^2}{\sigma^2}$$

Can prove it for a fixed $\boldsymbol{z}$

Summary

- provable (non)convex optimization with generic coefficients via local search
- the main challenge in nonconvex optimization is concentration
- for realistic data models need generic chaining

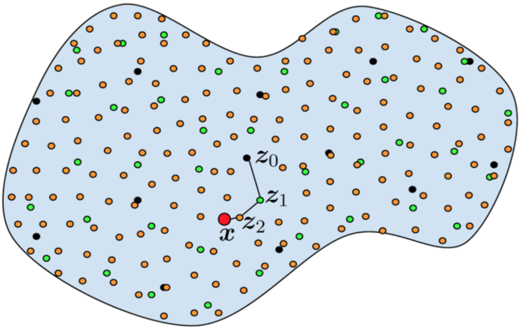
Summary

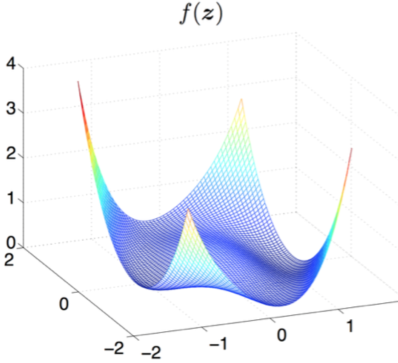
- provable (non)convex optimization with generic coefficients via local search
- the main challenge in nonconvex optimization is concentration
- for realistic data models need generic chaining

References:

- Phase retrieval
  - Phase retrieval via Wirtinger flow: Theory and algorithms E. J. Candes, X. Li, and M. Soltanolkotabi
  - Algorithms and theory for clustering and non-convex quadratic programming. M. Soltanolkotabi 2014.
  - Experimental robustness of Fourier Ptychography phase retrieval algorithms 2015 (in collaboration with the computational imaging lab at UC Berkeley)
- Low-rank matrix recovery
  - Low-rank Solutions of Linear Matrix Equations via Procrustes Flow. S. Tu, R. Boczar, M. Soltanolkotabi, and B. Recht 2015.
- Sharp time-data tradeoffs for (non)convex projected gradients
  - Sharp Time–Data Tradeoffs for Linear Inverse Problems. S. Oymak, B. Recht and M. Soltanolkotabi
  - Fast and reliable parameter estimation from nonlinear observations. S. Oymak and M. Soltanolkotabi.

Generic chaining meets nonconvex optimization





Thanks!

Just follow the gradient?