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 Candès
Linear inverse problems

\[ y = Ax + w \]
Compressed Sensing

Compressed sensing is a method that allows for the reconstruction of linear measurements from a structured signal.

[Photo credit Candes, Romberg, Tao 2006]

linear measurements from a structured signal.
Recommender systems

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>4</td>
<td>?</td>
<td>?</td>
<td>4</td>
<td>?</td>
</tr>
<tr>
<td>?</td>
<td>?</td>
<td>?</td>
<td>4</td>
<td>4</td>
<td>?</td>
</tr>
<tr>
<td>?</td>
<td>4</td>
<td>4</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Movie Features

User Features
The “power” of convex programing

Exciting research over the last decade demonstrating the effectiveness of convex programming/greedy algorithms.
The “power” of convex programming

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 $\ell_1$ 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 $\ell_1$ 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!

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

- 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)
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

\[ y_r = |\langle a_r, x \rangle|^2 \quad r = 1, 2, \ldots, m \quad \Leftrightarrow \quad y = |Ax|^2 \]

Find a feasible point in the intersection of quadratic equations

\[ y_r = x^* A_r x \quad \text{for} \quad r = 1, 2, \ldots, m. \]
Solving quadratic equations

\[ y_r = \left| \langle a_r, x \rangle \right|^2 \quad r = 1, 2, \ldots, m \quad \Leftrightarrow \quad y = |Ax|^2 \]

Find a feasible point in the intersection of quadratic equations

\[ y_r = x^* A_r x \quad \text{for} \quad r = 1, 2, \ldots, m. \]
Solving quadratic equations

\[ y_r = |\langle a_r, x \rangle|^2 \quad r = 1, 2, \ldots, m \quad \Leftrightarrow \quad y = |Ax|^2 \]

Find a feasible point in the intersection of quadratic equations

\[ y_r = x^* A_r x \quad \text{for} \quad r = 1, 2, \ldots, m. \]

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

- Detectors only record intensities of diffracted rays \((\text{magnitude measurements only!})\)

- 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!)

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

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 $\mathbf{x} \in \mathbb{C}^n$

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

$\mathbf{f}_k^*$ is $k$th row of the DFT matrix.

Phase retrieval is impossible, inherent ambiguity.
Phase retrieval

Measurements of the form

\[ y_{k \ell} = |f_k^* D_{\ell}^* x|^2 \quad k = 1, 2, \ldots, n \quad \text{and} \quad \ell = 1, 2, \ldots, L. \]

\[ y_r = |a_r^* x|^2 \quad r = 1, 2, \ldots, m. \]

Here \( m = nL \) and \( a_r = D_{\ell} f_k \), \( r \) represents \((k, \ell)\).
Phase retrieval

Measurements of the form

\[ y_{k\ell} = |f_k^* D_\ell^* x|^2 \quad k = 1, 2, \ldots, n \quad \text{and} \quad \ell = 1, 2, \ldots, L. \]
\[ y_r = |a_r^* x|^2 \quad r = 1, 2, \ldots, m. \]

Here \( m = nL \) and \( a_r = D_\ell f_k, r \) represents \((k, \ell)\)
Phase Retrieval by non-convex optimization

Let \( A = [a_1, a_2, \ldots, a_m] \)

\[
\min_{z \in \mathbb{C}^n} f(z) := \frac{1}{2m} \left\| y - |A^*z|^2 \right\|_{\ell_2}^2
\]

\[
= \frac{1}{2m} \sum_{r=1}^{m} (y_r - |a_r^*z|^2)^2
\]

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

- **Pro**: operates over vectors much less intensive!
- **Con**: Non-convex!
Algorithm 1 Wirtinger Flow (WF)

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

Initialization (WF-INIT):
Set $\tilde{z}_0$ to be the eigenvector corresponding to the largest eigenvalue of
$$ Y = \frac{1}{m} \sum_{r=1}^{m} y_r a_r a^*_r. $$

Set $z_0 = \left( \sqrt{\frac{1}{m} \sum_{r=1}^{m} y_r} \right) \tilde{z}_0$.

Iterations:
\[ \text{for } \tau = 0 \text{ to } t - 1 \text{ do} \]
\[ \text{Set} \]
$$ z_{\tau+1} = z_\tau - \frac{\mu_{\tau+1}}{\|z_0\|_2^2} \left( \frac{1}{m} \sum_{r=1}^{m} (|a_r^* z|^2 - y_r) (a_r a_r^*) z \right) := z_\tau - \frac{\mu_{\tau+1}}{\|z_0\|_2^2} \nabla f(z_\tau). \]

end for

Output: $\hat{x} = z_t$. 
Exact Phase Retrieval by WF (Gaussian Model)

For a vector \( z \in \mathbb{C}^n \)

\[
\text{dist}(z, x) = \min_{\phi \in [0, 2\pi]} \| z - e^{i\phi} x \|_{\ell^2}.
\]

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

Assume \( m \gtrsim n \). Using \( 0 \leq \mu \leq \mu_0/n \), with high probability

- **Initialization:**
  \[
  \text{dist}(z_0, x) \leq \sqrt{\frac{5}{6}} \| x \|_{\ell^2}.
  \]

- **After \( t \) iterations:**
  \[
  \text{dist}(z_t, x) \leq e^{-c\mu t} \cdot \text{dist}(z_0, x) \leq \sqrt{\frac{5}{6}} e^{-c\mu t} \| x \|_{\ell^2}.
  \]

[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 = $o(nL \log n)$.

**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 = $o(nL \log n)$.

**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.
WF after 500 iterations: Time=221.7697 sec, Relerr=2.5410 \times 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 $xx^*$
A real experiment

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

SDP based methods (PhaseLift, PhaseCut, ...) require $xx^*$
$(320 \times 1280)^2 \times 8$ Bytes $\approx 0.16777$ Tera Bytes
A real experiment

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

SDP based methods (PhaseLift, PhaseCut, ...) require $xx^*$

$(320 \times 1280)^2 \times 8 \text{ Bytes} \approx 0.16777 \text{ Tera Bytes}$

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

\[ f(z) = \frac{1}{2m} \sum_{r=1}^{m} \left( y_r - |a_r^* z|^2 \right)^2 \]

and

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

\[ f(z) = \frac{1}{2m} \sum_{r=1}^{m} (y_r - |a_r^* z|^2)^2 \]

and

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

Showing \( f(z) \approx F(z) \) not sufficient. Need to show this in terms of higher order derivatives.
Regularity Condition

Define

\[ P := \{ xe^{i\phi} : \phi \in [0, 2\pi] \}. \]

and

\[ E(\epsilon) := \{ z \in \mathbb{C}^n : \text{dist}(z, P) \leq \epsilon \}. \]

**Condition (Regularity Condition)**

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

\[ \text{Re} \left( \langle \nabla f(z), z - xe^{i\phi(z)} \rangle \right) \geq \frac{1}{\alpha} \text{dist}^2(z, x) + \frac{1}{\beta} \| \nabla f(z) \|_{\ell_2}^2. \]
Regularity condition leads to convergence

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

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

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

\[ z_+ = z - \mu \nabla f(z) \]

obeys

\[
\text{dist}^2(z_+, x) \leq \left( 1 - \frac{2\mu}{\alpha} \right) \text{dist}^2(z, x).
\]

\[
\left\| z_+ - x e^{i\phi(z_+)} \right\|_{\ell_2}^2 \leq \left\| z_+ - x e^{i\phi(z)} \right\|_{\ell_2}^2 = \left\| z - x e^{i\phi(z)} - \mu \nabla f(z) \right\|_{\ell_2}^2
\]

\[
= \left\| z - x e^{i\phi(z)} \right\|_{\ell_2}^2 - 2\mu \text{Re} \left( \langle \nabla f(z), (z - x e^{i\phi(z)}) \rangle \right) + \mu^2 \| \nabla f(z) \|_{\ell_2}^2
\]

\[
\leq \left\| z - x e^{i\phi(z)} \right\|_{\ell_2}^2 - 2\mu \left( \frac{1}{\alpha} \left\| z - x e^{i\phi(z)} \right\|_{\ell_2}^2 + \frac{1}{\beta} \| \nabla f(z) \|_{\ell_2}^2 \right) + \mu^2 \| \nabla f(z) \|_{\ell_2}^2
\]

\[
= \left( 1 - \frac{2\mu}{\alpha} \right) \left\| z - x e^{i\phi(z)} \right\|_{\ell_2}^2 + \mu \left( \mu - \frac{2}{\beta} \right) \| \nabla f(z) \|_{\ell_2}^2
\]

\[
\leq \left( 1 - \frac{2\mu}{\alpha} \right) \left\| z - x e^{i\phi(z)} \right\|_{\ell_2}^2,
\]
Regularity condition leads to convergence

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

\[
\Re \left( \langle \nabla f(z), z - xe^{i\phi(z)} \rangle \right) \geq \frac{1}{\alpha} \dist^2(z, x) + \frac{1}{\beta} \|\nabla f(z)\|_{\ell_2}^2.
\]

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

\[
z_+ = z - \mu \nabla f(z)
\]

obeys

\[
\dist^2(z_+, x) \leq \left( 1 - \frac{2\mu}{\alpha} \right) \dist^2(z, x).
\]
How do we prove the regularity condition?

\[
\Re \left( \langle \nabla f(z), z - xe^{i\phi(z)} \rangle \right) \geq \frac{1}{\alpha} \text{dist}^2(z, x) + \frac{1}{\beta} \| \nabla f(z) \|^2_{\ell_2}.
\]

and

\[
\nabla f(z) = \frac{1}{m} \sum_{r=1}^{m} \left( |a^*_rz|^2 - y_r \right) (a^*_rz)a_r.
\]

**Condition (Local Curvature Condition)**

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

**Condition (Local Smoothness Condition)**

\[
\| \nabla f(z) \|^2_{\ell_2} \leq \beta \left( \frac{(1 - \delta)}{4} \text{dist}^2(z, x) + \frac{1}{10m} \sum_{r=1}^{m} |a^*_r(z - e^{i\phi(z)}x)|^4 \right).
\]
Story II:
Linear inverse problems
Structured signal recovery/Linear inverse problems

Optimization Problem

\[ \hat{x} = \arg \min_z \frac{1}{2} \sum_{i=1}^{m} (y_i - \langle a_i, z \rangle)^2 = \frac{1}{2} \| y - Az \|_{\ell_2}^2 \]

subject to \( f(z) \leq f(x) \)

\( f \) is a function that enforces structure

\[ y = Ax + w \]
Minimal data/measurements for exact recovery?

(Review)
\[ y = Ax, \ y \in \mathbb{R}^m, \ A \in \mathbb{R}^{m \times n}, \ \text{and} \ x \in \mathbb{R}^n \ \text{with} \ m < < n. \]

\[ \hat{x} = \arg\min_z \frac{1}{2} \| y - Az \|_2^2 \ \text{subject to} \ f(z) \leq f(x). \]

When is \( \hat{x} = x? \ m? \)
Why does $\ell_1$ work?
Why $\ell_1$ may not always work

$y = Ax$
Geometry

\[ C = \{ h : \| x + th \| \leq \| x \| \text{ for some } t > 0 \} \quad \text{cone of descent} \]

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

\[ C = \{ h : \|x + th\| \leq \|x\| \text{ for some } t > 0 \} \quad \text{cone of descent} \]

Exact recovery if \( C \cap \text{null}(A) = \{0\} \)
Mean width

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

- finite set \( \mathcal{T} \)
Mean width

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

- finite set \( \mathcal{T} \)

\[ \omega^2(\mathcal{T}) = 2 \log |\mathcal{T}| \]
Mean width

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

- finite set \( \mathcal{T} \)

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

- \( \mathcal{C} \) cone of decent of \( \ell_1 \) norm at an \( s \)-sparse signal
Mean width

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

- **finite set** \(\mathcal{T}\)

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

- \(\mathcal{C}\) cone of decent of \(\ell_1\) 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

\[ m \geq m_0(f, x) := \omega^2(C_f(x)), \]

then

\[ \hat{x} = x \]

holds with high probability.

- \( f \) is convex and \( A \) 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 $\ell_1$ 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

\[ \hat{x} = \arg\min_z \frac{1}{2} \|y - Az\|_2^2 \quad \text{subject to} \quad f(z) \leq f(x). \]

Projection: \( P_K(z), K = \{ z \in \mathbb{R}^n : f(z) \leq f(x) \} \)
Projected Gradient Descent

\[ \hat{x} = \arg\min_z \frac{1}{2} \|y - Az\|_2^2 \quad \text{subject to} \quad f(z) \leq f(x). \]

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

- Start from \( z_0 = 0 \).
- Apply the update \( z_{\tau+1} = \mathcal{P}_\mathcal{K}\left(z_\tau + \mu A^T (y - Az_\tau)\right) \)

\[ f(z) = \|z\|_{\ell_1} \]

\[ f(z) = \|z\|_{\ell_1/2}^2 \]

\[ f(z) = \|z\|_{\ell_0} \]
Theory
A simple thought experiment

Reminder: $y = Ax + w$.

Algorithm:

- Start with $z_0 = 0$,
- Run $z_{	au+1} = P_K(z_{	au} + \mu A^T(y - Az_{	au}))$
A simple thought experiment

Reminder: \( \mathbf{y} = A\mathbf{x} + \mathbf{w} \).

Algorithm:

- Start with \( z_0 = 0 \),
- Run \( z_{\tau+1} = \mathcal{P}_K (z_{\tau} + \mu A^T (\mathbf{y} - A z_{\tau})) \)

Forget about the projection: \( z_{\tau+1} = z_{\tau} + \mu A^T (\mathbf{y} - A z_{\tau}) \)
A simple thought experiment

Reminder: \( y = Ax + w \).

Algorithm:

- Start with \( z_0 = 0 \),
- Run \( z_{\tau + 1} = \mathcal{P}_K (z_\tau + \mu A^T (y - Az_\tau)) \)

Forget about the projection: \( z_{\tau + 1} = z_\tau + \mu A^T (y - Az_\tau) \)

\[
z_{\tau + 1} - x = (I - \mu A^* A)(z_\tau - x) - \mu A^* w.
\]
A simple thought experiment

Reminder: $y = Ax + w$.

Algorithm:

- Start with $z_0 = 0$,
- Run $z_{\tau+1} = \mathcal{P}_\mathcal{K} \left( z_\tau + \mu A^T (y - Az_\tau) \right)$

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

$$z_{\tau+1} - x = (I - \mu A^* A)(z_\tau - x) - \mu A^* w.$$
Definition (Descent Cone)

\[ C_f(x) = \{ h : f(x + \tau h) \leq f(x) \} \].
A deterministic result

Algorithm:
- Start with $z_0 = 0$,
- Run $z_{\tau+1} = \mathcal{P}_\mathcal{K}(z_{\tau} + \mu A^T(y - Az_{\tau}))$
A deterministic result

Algorithm:
- Start with $z_0 = 0$,
- Run $z_{\tau+1} = \mathcal{P}_\mathcal{K} \left( z_\tau + \mu A^T (y - Az_\tau) \right)$

Theorem (Oymak, Soltanolkotabi, and Recht 2016)

For any $A$, and any convex $f$

$$
\|z_\tau - x\|_{\ell_2} \leq (\rho(\mu))^\tau \|x\|_{\ell_2} + \frac{1 - (\rho(\mu))^\tau}{1 - \rho(\mu)} \xi(\mu) \|w\|_{\ell_2}.
$$

- $\rho(\mu)$ is the convergence rate
  $$
  \rho(\mu) := \rho(\mu, A, f, x) = \sup_{u,v \in C_f(x) \cap B^n} u^* (I - \mu A^* A) v,
  $$

- $\xi(\mu)(A)$ is the noise amplification factor
  $$
  \xi(\mu)(A) := \xi(\mu, A, f, x, w) = \mu \cdot \sup_{v \in C_f(x) \cap B^n} v^* A^* \frac{w}{\|w\|_{\ell_2}}.
  $$
A deterministic result

Algorithm:
- Start with $z_0 = 0$,
- Run $z_{\tau+1} = \mathcal{P}_\mathcal{K} \left( z_\tau + \mu A^T (y - A z_\tau) \right)$

Theorem (Oymak, Soltanolkotabi, and Recht 2016)

For any $A$, and any convex $f$

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

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

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

- for non-convex $f$ just replace $\rho$ with $2\rho$!
For convex $f$ traditional literature will tell you $1/\tau$ 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 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

\[ y = Ax + w \]

Algorithm:
- Start with \( z_0 = 0 \),
- Run \( z_{\tau + 1} = \mathcal{P}_K \left( z_\tau + \mu A^T (y - Az_\tau) \right) \)

**Theorem (Oymak, Soltanolkotabi, and Recht 2016)**

For Gaussian \( A \), \( \mu = 1/m \), with high probability
\[
\| z_\tau - x \|_2 \leq \left( \sqrt{\frac{8}{\kappa^2 f}} \right)^\tau \| z_0 - x \|_2 + \kappa f \sqrt{\frac{\pi}{2}} \sqrt{\frac{1}{m_0}} \| w \|_2
\]

Sample complexity \( m \geq \frac{8}{\kappa^2 f^2} m_0 \)

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 greedy step size

\( y = Ax + w \)

**Algorithm:**
- Start with \( z_0 = 0 \),
- Run \( z_{\tau+1} = \mathcal{P}_K (z_\tau + \mu A^T (y - Az_\tau)) \)

**Theorem (Oymak, Soltanolkotabi, and Recht 2016)**

For Gaussian \( A \), \( \mu = 1/m \), with high probability

\[
\| z_\tau - x \|_{\ell_2} \leq \left( \sqrt{8 \kappa_f^2 \frac{m_0}{m}} \right)^\tau \| z_0 - x \|_{\ell_2} + \kappa_f \sqrt{\frac{\pi}{2}} \frac{\sqrt{m_0}}{m} \| w \|_{\ell_2}
\]

- 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

\( y = Ax + w \) Algorithm:

- Start with \( z_0 = 0 \),
- Run \( z_{\tau+1} = P_K \left( z_\tau + \mu A^T (y - Az_\tau) \right) \)
Gaussian measurements with conservative step size

\[ y = Ax + w \]

Algorithm:
- Start with \( z_0 = 0 \),
- Run \( z_{\tau+1} = P_K (z_{\tau} + \mu A^T (y - A z_{\tau})) \)

Theorem (2016)

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

\[
\| z_{\tau} - x \|_{\ell_2} \leq \left( 1 - \frac{0.3}{m + n} (\sqrt{m} - \sqrt{m_0})^2 \right)^{\tau} \cdot \| x \|_{\ell_2} + \frac{3.5}{(1 - \sqrt{m_0/m})^2} \frac{\sqrt{m_0}}{m} \| w \|_{\ell_2}.
\]

- Sample complexity \( m \geq m_0 \)
- 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

\[ y = g(Ax). \]

How can we recover \( x \) from these measurements?

Theorem (Oymak and Soltanolkotabi 2016)

For Gaussian \( A \) and \( \mu = 1/m \), with high probability

\[ \| z_\tau - \mu x \|_2 \leq \sqrt{8 \kappa^2 f m_0 m} \| z_0 - \mu x \|_2 + 2 \sigma \sqrt{m_0 m}, \]

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

Interpretation

\( g(Ax) \approx \mu Ax + \sigma w \).

Nonlinear Observations

Nonlinear observe from a structured signal

\[ y = g(Ax). \]

How can we recover \( x \) from these measurements?

Pretend the model was linear and run the same algorithm

- Start with \( z_0 = 0 \),
- Run \( z_{\tau+1} = P_K \left( z_\tau + \mu A^T (y - Az_\tau) \right) \)

Theorem (Oymak and Soltanolkotabi 2016)

For Gaussian \( A \) and \( \mu = 1/m \), with high probability

\[ \| z_\tau - \mu x \|_{\ell^2} \leq \left( \sqrt{8 \kappa^2 f m_0 m} \right) \| z_0 - \mu x \|_{\ell^2} + 2 \sigma \sqrt{m_0 m}, \]

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

Interpretation

\( g(Ax) \approx \mu Ax + \sigma w \).

Nonlinear Observations

Nonlinear observe from a structured signal

\[ y = g(Ax). \]

How can we recover \( x \) from these measurements?

Pretend the model was linear and run the same algorithm

- Start with \( z_0 = 0 \),
- Run \( z_{\tau+1} = \mathcal{P}_\mathcal{K} \left( z_\tau + \mu A^T(y - Az_\tau) \right) \)

\[ \text{Theorem (Oymak and Soltanolkotabi 2016)} \]

For Gaussian \( A \) and \( \mu = 1/m \), with high probability

\[ \|z_\tau - \mu x\|_2 \leq \left( \sqrt{8\kappa_f^2 \frac{m_0}{m}} \right) \|z_0 - \mu x\|_2 + 2\sigma \sqrt{\frac{m_0}{m}}, \]

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

Interpretation

\[ g(Ax) \approx \mu Ax + \sigma w. \]

Non-Gaussians
### Definition (Subsampled Orthogonal with Random Sign (SORS) matrices)

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

\[
F^* F = I \quad \text{and} \quad \max_{i,j} |F_{ij}| \leq \frac{\Delta}{\sqrt{n}}.
\]

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

\( y = Ax + w \) Algorithm:

- Start with \( z_0 = 0 \),
- Run \( z_{\tau+1} = P_{\mathcal{K}} \left( z_\tau + \mu A^T (y - Az_\tau) \right) \)

Theorem (Oymak, Recht and Soltanolkotabi 2016)

- \( A \) is a SORS matrix
- \( m > c \Delta \cdot (\log n) \cdot m^0 \), with high probability

\[ \| z_\tau - x \|_2 \leq (\frac{c \Delta m^0 m \log 4}{n})^{\frac{\tau^2}{2}} \| x \|_2. \]

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

For sparse matrices see [Bourgain, Dirksen, Nelson 2014]

Can we get down to \( c \Delta (\log n) (\log m^0) \)? Maybe by [Haviv and Regev 15]?
ORS results

\[ y = Ax + w \]

Algorithm:

- Start with \( z_0 = 0 \),
- Run \( z_{\tau+1} = P_K (z_{\tau} + \mu A^T (y - Az_{\tau})) \)

---

**Theorem (Oymak, Recht and Soltanolkotabi 2016)**

\( A \) is a SORS matrix

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

with high probability

\[ \| z_\tau - x \|_{\ell_2} \leq \left( c_\Delta \frac{m_0}{m} \log^4 n \right)^{\frac{\tau}{2}} \| x \|_{\ell_2}. \]
SORS results

\( \mathbf{y} = A\mathbf{x} + \mathbf{w} \) Algorithm:
- Start with \( z_0 = 0 \),
- Run \( z_{\tau+1} = \mathcal{P}_{\mathcal{K}} \left( z_{\tau} + \mu A^T (\mathbf{y} - Az_{\tau}) \right) \)

Theorem (Oymak, Recht and Soltanolkotabi 2016)

\( A \) is a SORS matrix

\[
m > c_{\Delta} \cdot (\log n)^4 \cdot m_0,\]

with high probability

\[
\| z_{\tau} - \mathbf{x} \|_{\ell_2} \leq \left( c_{\Delta} \frac{m_0}{m} \log^4 n \right)^{\frac{\tau}{2}} \| \mathbf{x} \|_{\ell_2}.
\]

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

For sparse matrices see [Bourgain, Dirksen, Nelson 2014]
Can we get down to \( c_{\Delta} \Delta (\log n)(\log m_0)^2 \)? Maybe by [Haviv and Regev 15]?
**SORs results**

\[ \mathbf{y} = A\mathbf{x} + \mathbf{w} \]

**Algorithm:**
- Start with \( z_0 = 0 \),
- Run \( z_{\tau+1} = \mathcal{P}_K \left( z_{\tau} + \mu A^T (\mathbf{y} - A z_{\tau}) \right) \)

---

**Theorem (Oymak, Recht and Soltanolkotabi 2016)**

\( A \) is a SORS matrix

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

with high probability

\[ \| z_{\tau} - \mathbf{x} \|_2 \leq \left( c_\Delta \frac{m_0}{m} \log^4 n \right)^{\tau/2} \| \mathbf{x} \|_2. \]

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

\[ y = Ax + w \]

Algorithm:
- Start with \( z_0 = 0 \),
- Run \( z_{\tau+1} = \mathcal{P}_K \left( z_\tau + \mu A^T (y - Az_\tau) \right) \)

Theorem (Oymak, Recht and Soltanolkotabi 2016)

\( A \) is a SORS matrix

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

with high probability

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

- 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

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

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

Then for all $u \in \mathcal{C}$
- with prob. at least $1 - e^{-\frac{\eta^2}{2}}$
  \[ \frac{\|Au\|_{\ell_2}}{\|u\|_{\ell_2}} \geq b_m - (\omega(\mathcal{C} \cap \mathbb{S}^{n-1}) + \eta) \]
- with prob. at least $1 - e^{-\frac{\eta^2}{2}}$
  \[ \frac{\|Au\|_{\ell_2}}{\|u\|_{\ell_2}} \leq b_m + (\omega(\mathcal{C} \cap \mathbb{S}^{n-1}) + \eta) \]
Gordon’s lemma

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

Then for all $u \in C$

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

$$\frac{\|Au\|_{\ell_2}}{\|u\|_{\ell_2}} \geq b_m - (\omega(C \cap S^{n-1}) + \eta)$$

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

$$\frac{\|Au\|_{\ell_2}}{\|u\|_{\ell_2}} \leq b_m + (\omega(C \cap S^{n-1}) + \eta)$$

Thus for all $u \in C$

$$-2\frac{\omega + \eta}{b_m} \|u\|_{\ell_2}^2 + \left(\frac{\omega + \eta}{b_m}\right)^2 \|u\|_{\ell_2}^2 \leq \frac{1}{b_m^2} \|Au\|_{\ell_2}^2 - \|u\|_{\ell_2}^2 \leq 2\frac{\omega + \eta}{b_m} \|u\|_{\ell_2}^2 + \left(\frac{\omega + \eta}{b_m}\right)^2$$
Bounding the convergence rate in the convex case

\[ \rho(\frac{1}{m}) = \sup_{u,v \in C_f(x) \cap S^{n-1}} u^T \left( I - \frac{1}{m} A^T A \right) v \]
Bounding the convergence rate in the convex case

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

Simple algebra

\[ u^T \left( I - \frac{1}{m} A^T A \right) v = 2 \left( \left\| \frac{u + v}{2} \right\|_{\ell_2}^2 - \frac{1}{m} \left\| A \left( \frac{u + v}{2} \right) \right\|_{\ell_2}^2 \right) \]

\[ + \frac{1}{2} \left( \frac{\left\| Au \right\|_{\ell_2}^2}{m} - \left\| u \right\|_{\ell_2}^2 \right) \]

\[ + \frac{1}{2} \left( \frac{\left\| Av \right\|_{\ell_2}^2}{m} - \left\| v \right\|_{\ell_2}^2 \right). \]
Gordon for SORS matrices?

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

\[ F \in \mathbb{R}^{n \times n} \text{ orthonormal matrix} \]

\[ F^* F = I \quad \text{and} \quad \max_{i,j} |F_{ij}| \leq \frac{\Delta}{\sqrt{n}}. \]

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

Gordon Holds for SORS matrices
Gordon for SORS matrices?

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

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

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

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

Gordon Holds for SORS matrices

Theorem (Oymak, Recht, and Soltanolkotabi 2016)

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

$$\sup_{x \in \mathcal{T}} \left| \|Ax\|_2^2 - \|x\|_2^2 \right| \leq \max(\delta, \delta^2) \cdot (\text{rad}(\mathcal{T}))^2,$$

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

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

This would settle universality conjecture in compressive sensing!
Open problem

- $\mathcal{C} \in \mathbb{R}^n$
- $A \in \mathbb{R}^{m \times n}$, sub-sampled Fourier multiplied random sign.
- $b_m = \mathbb{E}[\|g\|_2]$ with $g \in \mathbb{R}^m$ distributed as $\mathcal{N}(0, I_m)$

Conjecture

- with prob. at least $1 - ???$
  \[
  \inf_{u \in \mathcal{C}} \frac{\|Au\|_2}{\|u\|_2} \geq b_m - (\omega(\mathcal{C} \cap S^{n-1}) + \eta)
  \]

- with prob. at least $1 - ????$
  \[
  \sup_{u \in \mathcal{C}} \frac{\|Au\|_2}{\|u\|_2} \leq (b_m + (\omega(\mathcal{C} \cap S^{n-1}) + \eta)) \sqrt{\log n}????
  \]

This would settle universality conjecture in compressive sensing!
SORS matrices: from RIP to JL

- Restricted Isometry Property-RIP($s, \delta$)

\[
\sup_{x: \|x\|_{\ell_0} \leq s} \left| \|Fx\|_{\ell_2}^2 - \|x\|_{\ell_2}^2 \right| \leq \max(\delta, \delta^2) \|x\|_{\ell_2}^2
\]

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

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

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

Then $A = HD$ with $D$ diagonal sign pattern obeys

\[
\sup_{x \in \mathcal{T}} \left| \|Ax\|_{\ell_2}^2 - \|x\|_{\ell_2}^2 \right| \leq \max(\epsilon, \epsilon^2) \|x\|_{\ell_2}^2,
\]

with high probability.
From JL to Gordon?

\[ m \gtrsim \log |N| \delta^2 \log 4n \sim n^m \omega^2 (T) \delta^4 \log 4n \iff m \gtrsim \sqrt{n} \omega (T) \delta^2 \log 2n, \]
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 $|\mathcal{T}_\ell| = 2^{2^\ell}$
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 $|T_\ell| = 2^{2^\ell}$
- Different distortion levels $\delta_\ell = 2^{\ell/2} \frac{\delta}{\omega(T)}$

$$m \gtrsim \max_{\ell=1,2,\ldots,L} \frac{\log |T_\ell|}{\delta_\ell^2} \log^4 n = \max_{\ell=1,2,\ldots,L} \frac{\log 2^{2^\ell}}{2^\ell \frac{\delta_\ell^2}{\omega^2(T)}} \log^4 n = \frac{\omega^2(T)}{\delta^2} \sim \log^4 n \frac{\omega^2(T)}{\delta^2},$$
Is the overall distortion small enough?

After some work Discrete JL implies that with high probability

- For all \( \mathbf{v} \in \mathcal{T}_{\ell-1} \cup \mathcal{T}_\ell \cup (\mathcal{T}_{\ell-1} - \mathcal{T}_\ell) \),
  \[
  \|A \mathbf{v}\|_{\ell_2} \leq (1 + \delta_{\ell}) \|\mathbf{v}\|_{\ell_2}.
  \]

- For all \( \mathbf{v} \in \mathcal{T}_{\ell-1} \cup \mathcal{T}_\ell \cup (\mathcal{T}_{\ell-1} - \mathcal{T}_\ell) \),
  \[
  \|A \mathbf{v}\|_{\ell_2}^2 - \|\mathbf{v}\|_{\ell_2}^2 \leq \max(\delta_{\ell}, \delta_{\ell}^2) \cdot \|\mathbf{v}\|_{\ell_2}^2.
  \]

- For all \( \mathbf{u} \in \mathcal{T}_{\ell-1} \) and \( \mathbf{v} \in \mathcal{T}_\ell - \{\mathbf{u}\} := \{\mathbf{y} - \mathbf{u} : \mathbf{y} \in \mathcal{T}_\ell\} \),
  \[
  |\mathbf{u}^* A^* A \mathbf{v} - \mathbf{u}^* \mathbf{v}| \leq \max(\delta_{\ell}, \delta_{\ell}^2) \cdot \|\mathbf{u}\|_{\ell_2} \|\mathbf{v}\|_{\ell_2}.
  \]

where

\[
\delta_{\ell} = 2^{\ell/2} \frac{\delta}{\omega(\mathcal{T})}
\]
Is the overall distortion small enough?

We are interested in bounding $\|Ax\|_2^2 - \|x\|_2^2$ for all $x \in \mathcal{T}$. Define $\tilde{L} = \max\left(0, \left\lfloor 2 \log_2 \left(\frac{\omega(T)}{\delta}\right) \right\rfloor\right)$

$$\|Ax\|_2^2 - \|x\|_2^2 \leq |\|Az_{\tilde{L}}\|_2^2 - \|z_{\tilde{L}}\|_2^2|$$

$$+ |\|Ax\|_2^2 - \|Az_{\tilde{L}}\|_2^2| + |\|x\|_2^2 - \|z_{\tilde{L}}\|_2^2|$$

$$\leq \sum_{\ell=1}^{\tilde{L}} \left( |\|Az_{\ell}\|_2^2 - \|z_{\ell}\|_2^2| - |\|Az_{\ell-1}\|_2^2 - \|z_{\ell-1}\|_2^2| \right)$$

$$+ |\|Ax\|_2^2 - \|Az_{\tilde{L}}\|_2^2| + |\|x\|_2^2 - \|z_{\tilde{L}}\|_2^2|$$

$$+ |\|Az_0\|_2^2 - \|z_0\|_2^2|.$$
Is the overall distortion small enough? (First term)

For the first term

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

where $e_\ell = \text{dist}(x, T_\ell)$. Then

$$\sum_{\ell=1}^{\tilde{L}} \left( \left| \| A z_\ell \|_{\ell_2}^2 - \| z_\ell \|_{\ell_2}^2 \right| - \left| \| A z_{\ell-1} \|_{\ell_2}^2 - \| z_{\ell-1} \|_{\ell_2}^2 \right| \right) \leq 10 \frac{\delta}{\omega(T)} \left( \sum_{\ell=1}^{\tilde{L}} 2^{\ell/2} e_{\ell-1} \right)$$

$$= 10\sqrt{2} \frac{\delta}{\omega(T)} \left( \sum_{\ell=0}^{\tilde{L}-1} 2^{\ell/2} e_\ell \right)$$

$$= 10\sqrt{2} \frac{\delta}{\omega(T)} \gamma_2(T)$$

$$\leq c\delta.$$
Is the overall distortion small enough? (second term)

\[
\|A\mathbf{x}\|_2 - \|A\mathbf{z}_L\|_2 = \|A\mathbf{x}\|_2 - \|A\mathbf{z}_L\|_2 + \|A\mathbf{z}_L\|_2 - \|A\mathbf{z}_L\|_2
\]

\[
\leq \|A(x - z_L)\|_2 + \|A(z_L - z_L)\|_2
\]

\[
\leq \|A\| \|x - z_L\|_2 + \left\| \sum_{\ell=L+1}^{L} A(z_\ell - z_{\ell-1}) \right\|_2
\]

\[
\leq \left( \frac{1}{4} \frac{L}{2} \frac{\delta}{\omega(T)} + 1 \right) e_L + \sum_{\ell=L+1}^{L} \left\| A(z_\ell - z_{\ell-1}) \right\|_2
\]

\[
\leq \left( \frac{1}{4} \frac{L}{2} \frac{\delta}{\omega(T)} + 1 \right) e_L + \sum_{\ell=L+1}^{L} \left( 1 + 2^{\ell/2} \frac{\delta}{\omega(T)} \right) \|z_\ell - z_{\ell-1}\|_2
\]

\[
\leq \frac{5}{4} \frac{2^{L/2}}{\omega(T)} \delta e_L + \sum_{\ell=L+1}^{L} 2^{\ell/2+1} \frac{\delta}{\omega(T)} \|z_\ell - z_{\ell-1}\|_2
\]

\[
\leq \frac{5}{4} \frac{\delta}{\omega(T)} 2^{L/2} e_L + 4\sqrt{2} \frac{\delta}{\omega(T)} \sum_{\ell=L+1}^{L} 2^{(\ell-1)/2} e_{\ell-1}
\]

\[
\leq 4\sqrt{2} \frac{\delta}{\omega(T)} \left( \sum_{\ell=L}^{L} 2^{\ell/2} e_{\ell} \right)
\]

\[
\leq 4\sqrt{2} \frac{\delta}{\omega(T)} \gamma_2(T)
\]
Is the overall distortion small enough? (second term cont.)

\[
\left| \|Ax\|_{\ell_2}^2 - \|Az_{\tilde{L}}^2\|_{\ell_2} \right| \leq \left| \|Ax\|_{\ell_2} - \|Az_{\tilde{L}}\|_{\ell_2} \right| \|Ax\|_{\ell_2} + \|Az_{\tilde{L}}\|_{\ell_2} \\
\leq \left| \|Ax\|_{\ell_2} - \|Az_{\tilde{L}}\|_{\ell_2} \right|^2 + 2 \left| \|Ax\|_{\ell_2} - \|Az_{\tilde{L}}\|_{\ell_2} \right| \|Az_{\tilde{L}}\|_{\ell_2} \\
\leq 32 \frac{\delta^2}{\omega^2(\mathcal{T})} \gamma_2^2(\mathcal{T}) + 16 \sqrt{2} \frac{\delta}{\omega(\mathcal{T})} \gamma_2(\mathcal{T}) \\
\leq c\delta.
\]
Is the overall distortion small enough? (third term)

Easy

\[ | \| A z_0 \|_2^2 - \| z_0 \|_2^2 | \leq \max \left( \frac{\delta}{\omega(T)}, \left( \frac{\delta}{\omega(T)} \right)^2 \right) \| z_0 \|_2^2 \]

\[ \leq \max \left( \frac{\delta}{\omega(T)}, \left( \frac{\delta}{\omega(T)} \right)^2 \right). \]
Is the failure probability small enough?

After some work Discrete JL implies that with high probability

- For all $v \in \mathcal{T}_{\ell-1} \cup \mathcal{T}_\ell \cup (\mathcal{T}_{\ell-1} - \mathcal{T}_\ell)$,
  $$\|Av\|_{\ell^2} \leq (1 + \delta_\ell) \|v\|_{\ell^2}.$$

- For all $v \in \mathcal{T}_{\ell-1} \cup \mathcal{T}_\ell \cup (\mathcal{T}_{\ell-1} - \mathcal{T}_\ell)$,
  $$\|\|Av\|_{\ell^2}^2 - \|v\|_{\ell^2}^2\| \leq \max(\delta_\ell, \delta_\ell^2) \cdot \|v\|_{\ell^2}^2.$$

- For all $u \in \mathcal{T}_{\ell-1}$ and $v \in \mathcal{T}_\ell - \{u\} := \{y - u : y \in \mathcal{T}_\ell\}$,
  $$|u^* A^* Av - u^* v| \leq \max(\delta_\ell, \delta_\ell^2) \cdot \|u\|_{\ell^2} \|v\|_{\ell^2}.$$

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?

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

Before Thresholding:
\[ z + A^T(y - Az) \]

After Thresholding:
\[ P_K(z + A^T(y - Az)) \]

Conclusion
Intuitively better de-noiser should work with less data
Minimal number of data?

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

Before Thresholding: \( z_{\tau} + A^T (y - Az_{\tau}) \)

After Thresholding: \( \mathcal{P}_K (z_{\tau} + A^T (y - Az_{\tau})) \)
Minimal number of data?

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

Before Thresholding: \( z_\tau + A^T (y - A z_\tau) \)

After Thresholding: \( \mathcal{P}_K (z_\tau + A^T (y - A z_\tau)) \)

**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 $A$, and convex $K$ minimal number of data is

$$\max_{\sigma} \frac{\mathbb{E} \left\| P_K(x + \sigma z) - x \right\|_{\ell_2}^2}{\sigma^2} = m_0$$

For non-convex $K$

$$\max_{\sigma} \frac{\mathbb{E} \left\| P_K(x + \sigma z) - x \right\|_{\ell_2}^2}{\sigma^2} \leq 4m_0$$

$m_0 \propto \# \text{ params e.g. } (2s + 1) \log(n/s)$
Add Gaussian noise to parameter and then de-noise!

**Theorem (Oymak, Recht and Soltanolkotabi 2016)**

For Gaussian $A$, and convex $K$ minimal number of data is

$$
\max_{\sigma} \frac{\mathbb{E} \left\| P_K(x + \sigma z) - x \right\|_{\ell_2}^2}{\sigma^2} = m_0
$$

For non-convex $K$

$$
\max_{\sigma} \frac{\mathbb{E} \left\| P_K(x + \sigma z) - x \right\|_{\ell_2}^2}{\sigma^2} \leq 4m_0
$$

$m_0 \propto \# \text{ params e.g. } (2s + 1) \log(n/s)$ 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:  After Thresholding:  Original signal $x$:

Relerr=0.70687  Relerr=0.57926
Iteration 2:

Before Thresholding:  

After Thresholding:  

Original signal $x$:  

Relerr = 0.54722  

Relerr = 0.45153
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 $x$: 
Relerr
Iteration 20:

Before Thresholding:  
After Thresholding:  
Original signal $x$:

Relerr=0.04864  
Relerr=0.061786
Iteration 40:

Before Thresholding:

After Thresholding:

Original signal $x$:

Relerr=0.027613

Relerr=0.031075
Iteration 100:

Before Thresholding:  

After Thresholding:  

Original signal $x$:

$\text{Relerr}=0.015708$  

$\text{Relerr}=0.015732$
Iteration 200:

Before Thresholding:  After Thresholding:  Original signal $x$:

Relerr=0.015603  Relerr=0.015618
Open Problem

Not very precise conjecture

Conjecture

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

\[
\|S(u) - S(v)\|_{\ell_2} \leq L \|u - v\|_{\ell_2}.
\]

show that for all \( z \in \text{Range}(S) \)

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

as long as

\[
m \geq \max_{\sigma} \frac{\mathbb{E} \|S(x + \sigma z) - x\|_{\ell_2}^2}{\sigma^2}
\]

Can prove it for a fixed \( z \)
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

References:
- Phase retrieval
  - Phase retrieval via Wirtinger flow: Theory and algorithms E. J. Candes, X. Li, and M. Soltanolkotabi
- Experimental robustness of Fourier Ptychography phase retrieval algorithms 2015 (in collaboration with the computational imaging lab at UC Berkeley)
- Low-rank matrix recovery
- 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.
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
  - Experimental robustness of Fourier Ptychography phase retrieval algorithms 2015 (in collaboration with the computational imaging lab at UC Berkeley)
- Low-rank matrix recovery
- 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?

\[ f(z) \]

\[ -f(z) \]