Optimization methods for inverse problems:
The times they are a-changin’

Carlos Fernandez-Granda
www.cims.nyu.edu/~cfgranda

10/22/2019
Separable nonlinear inverse problems

Machine learning for inverse problems
Separable nonlinear inverse problems

Machine learning for inverse problems
Acknowledgements

Joint work with Brett Bernstein, Sheng Liu and Chrysa Papadaniil

Project supported by NSF award DMS-1616340
Separable nonlinear (SNL) inverse problems

We consider phenomena governed by known nonlinear function $\phi_t$

Combination between sources or components is linear

$$f(t) := \sum_{i=1}^{k} c_i \phi_t(\theta_i) \quad (1)$$

Aim: estimate parameters $\theta_1, \ldots, \theta_k \in \mathbb{R}^d$ from $n$ samples

$$y := \begin{bmatrix} f(s_1) \\ \vdots \\ f(s_n) \end{bmatrix} = \sum_{i=1}^{k} c_i \vec{\phi}(\theta_i)$$
Spectral super-resolution

Classical problem in signal processing

Parameters encode frequencies of sinusoids

\[ \varphi_t(\theta_1) \]

\[ \varphi_t(\theta_2) \]

\[ \varphi_t(\theta_3) \]
Spectral super-resolution (data)

\[ y = \phi(\theta_1) + 2\phi(\theta_2) + 0.5\phi(\theta_3) \]
Deconvolution

Popular model in imaging and geophysics

Parameters encode spike locations

\[ \varphi_t(\theta_1) \]

\[ \varphi_t(\theta_2) \]

\[ \varphi_t(\theta_3) \]
Deconvolution (data)

\[ y = \vec{\phi}(\theta_1) + 2 \vec{\phi}(\theta_2) + 0.5 \vec{\phi}(\theta_3) \]
Heat source localization

Parameters encode source locations

Nonlinear function is obtained by solving the heat equation
Heat source localization (data)

\[ y = \vec{\phi}(\theta_1) + 2 \vec{\phi}(\theta_2) + 0.5 \vec{\phi}(\theta_3) \]
Electroencephalography

Parameters encode locations of brain activity
\vec{\phi}(\theta_1)
$\vec{\phi}(\theta_2)$
\[ y = \vec{\phi}(\theta_1) + \vec{\phi}(\theta_2) + \vec{\phi}(\theta_3) \]
Methods to tackle SNL problems

- Nonlinear least-squares solved by descent methods
  *Drawback:* local minima

- Prony-based / Finite-rate of innovation
  *Drawback:* challenging to apply beyond super-resolution

- Reformulate as sparse-recovery problem
  *Drawback:* very slow

- Learning-based methods
  *Drawback:* we don’t understand what’s going on
Linearization

Linearize problem by lifting to a higher-dimensional space

True parameters: $\theta_{T_1}, \ldots, \theta_{T_k}$

Grid of parameters: $\theta_1, \ldots, \theta_N$, $N >> n$

\[
y = \left[ \phi(\theta_1) \quad \cdots \quad \phi(\theta_{T_1}) \quad \cdots \quad \phi(\theta_{T_k}) \quad \cdots \quad \phi(\theta_N) \right] \begin{bmatrix} 0 \\ \cdots \\ c(1) \\ \cdots \\ c(s) \\ 0 \end{bmatrix} = \sum_{j=1}^{k} c(j) \phi(\theta_{T_j})
\]
Sparse Recovery for SNL Problems

Find a sparse $\tilde{c}$ such that

$$y = \Phi_{\text{grid}} \tilde{c}$$

Underdetermined linear inverse problem with sparsity prior
Popular approach: $\ell_1$-norm minimization

\[
\begin{align*}
\text{minimize} & & \|\tilde{c}\|_1 \\
\text{subject to} & & \Phi_{\text{grid}} \tilde{c} = y
\end{align*}
\]
Popular approach: $\ell_1$-norm minimization

- Deconvolution:
  *Deconvolution with the $\ell_1$ norm*, Taylor et al (1979)

- EEG:

- Direction-of-arrival in radar / sonar:
  *A sparse signal reconstruction perspective for source localization with sensor arrays*, Malioutov et al (2005)

- and many, many others...
Main question

Under what conditions can SNL problems be solved by $\ell_1$-norm minimization?
Under what conditions can SNL problems be solved by $\ell_1$-norm minimization?

*Wait, isn’t this just compressed sensing?*
Compressed sensing

Recover \( s \)-sparse vector \( x \) of dimension \( m \) from \( n < m \) measurements

\[
y = Ax
\]

Key assumption: \( A \) is random, and hence satisfies restricted-isometry properties with high probability
Restricted isometry property (Candès, Tao 2006)

An $m \times n$ matrix $A$ satisfies the restricted isometry property (RIP) if there exists $0 < \kappa < 1$ such that for any $s$-sparse vector $x$

\[
(1 - \kappa) \|x\|_2 \leq \|Ax\|_2 \leq (1 + \kappa) \|x\|_2
\]
Restricted isometry property (Candès, Tao 2006)

An $m \times n$ matrix $A$ satisfies the restricted isometry property (RIP) if there exists $0 < \kappa < 1$ such that for any $s$-sparse vector $x$

$$(1 - \kappa) \|x\|_2 \leq \|Ax\|_2 \leq (1 + \kappa) \|x\|_2$$

2s-RIP implies that for any $s$-sparse signals $x_1, x_2$

$$\|Ax_2 - Ax_1\|_2$$
Restricted isometry property (Candès, Tao 2006)

An \( m \times n \) matrix \( A \) satisfies the restricted isometry property (RIP) if there exists \( 0 < \kappa < 1 \) such that for any \( s \)-sparse vector \( x \)

\[
(1 - \kappa) \|x\|_2 \leq \|Ax\|_2 \leq (1 + \kappa) \|x\|_2
\]

2s-RIP implies that for any \( s \)-sparse signals \( x_1, x_2 \)

\[
\|Ax_2 - Ax_1\|_2 = \|A(x_2 - x_1)\|_2
\]
Restricted isometry property (Candès, Tao 2006)

An $m \times n$ matrix $A$ satisfies the restricted isometry property (RIP) if there exists $0 < \kappa < 1$ such that for any $s$-sparse vector $x$

$$(1 - \kappa) \|x\|_2 \leq \|Ax\|_2 \leq (1 + \kappa) \|x\|_2$$

2s-RIP implies that for any $s$-sparse signals $x_1, x_2$

$$\|Ax_2 - Ax_1\|_2 = \|A(x_2 - x_1)\|_2 \geq (1 - \kappa) \|x_2 - x_1\|_2$$
Columns of randomized matrix

$A_{75}$

$A_{150}$

$A_{225}$
Inter-column correlations

\[ |A_{75}^T A_i| \]
\[ |A_{150}^T A_i| \]
\[ |A_{225}^T A_i| \]
Separable nonlinear problems

Does RIP hold? Are all columns uncorrelated?
Correlations for spectral super-resolution

\[ |\rho_{\theta_1}(\eta)| \]

\[ \theta_1 \]

\[ |\rho_{\theta_2}(\eta)| \]

\[ \theta_2 \]

\[ |\rho_{\theta_3}(\eta)| \]

\[ \theta_3 \]
Correlations for deconvolution

\[ |\rho_{\theta_1}(\eta)| \]

\[ |\rho_{\theta_2}(\eta)| \]

\[ |\rho_{\theta_3}(\eta)| \]
Correlations for heat-source localization

\[ |\rho_{\theta_1}(\eta)| \]
\[ |\rho_{\theta_2}(\eta)| \]
\[ |\rho_{\theta_3}(\eta)| \]
Beyond sparsity

Due to high local correlations sparsity is not enough

Some sparse signals are impossible to estimate

But methods *work* in practice

**Goal:** Theory of sparse estimation relevant to SNL problems
Common property: Correlation decay
Minimum separation in parameter space

The minimum separation $\Delta$ of $\theta_1, \ldots, \theta_k$ equals

$$\Delta = \min_{i \neq j} |\theta_i - \theta_j|$$

A large enough minimum separation ensures that columns corresponding to active parameters are uncorrelated.

Empirical observation: Recovery is exact if $\Delta$ is large enough.
Spectral super-resolution
Deconvolution
Heat-source localization

![Graph showing Heat-source localization with Success and Failure axes and Δcorr on the x-axis. Success points are represented by blue diamonds, and Failure points by red dots.]
EEG

Success

Failure

Distance (cm)
Analysis of $\ell_1$-norm minimization

▶ **Aim:** Prove that if $Ax = y$ where $A$ has correlation decay and $x$ is well separated, then the solution to

$$\text{minimize} \quad \|x'\|_1$$

subject to $Ax' = y$

equals $x$
Analysis of $\ell_1$-norm minimization

- **Aim:** Prove that if $Ax = y$ where $A$ has correlation decay and $x$ is well separated, then the solution to

  $$\begin{align*}
  \text{minimize} & \quad \|x'\|_1 \\
  \text{subject to} & \quad Ax' = y
  \end{align*}$$

  equals $x$

- **Strategy:** Build dual certificate associated to an arbitrary well-separated $x$
The subgradient of \( f : \mathbb{R}^n \to \mathbb{R} \) at \( x \in \mathbb{R}^n \) is a vector \( g \in \mathbb{R}^n \) such that

\[
f(y) \geq f(x) + g^T (y - x), \quad \text{for all } y \in \mathbb{R}^n
\]

The set of all subgradients at \( x \) is called the subdifferential.
Subdifferential of $\ell_1$ norm

$g$ is a subgradient of the $\ell_1$ norm at $x \in \mathbb{R}^n$ if and only if

$$g[i] = \text{sign}(x[i]) \quad \text{if } x[i] \neq 0$$

$$|g[i]| \leq 1 \quad \text{if } x[i] = 0$$
Subdifferential of $\ell_1$ norm
Subdifferential of $\ell_1$ norm
Subdifferential of $\ell_1$ norm
Dual certificate

\( v \in \mathbb{R}^m \) is a dual certificate associated to \( x \) if

\[
q := A^T v
\]

satisfies

\[
q_i = \text{sign}(x_i) \quad \text{if } x_i \neq 0
\]

\[
|q_i| < 1 \quad \text{if } x_i = 0
\]
Dual certificate

$v \in \mathbb{R}^m$ is a dual certificate associated to $x$ if

$$q := A^T v$$

satisfies

$$q_i = \text{sign} (x_i) \quad \text{if } x_i \neq 0$$

$$|q_i| < 1 \quad \text{if } x_i = 0$$

$q$ is a subgradient of the $\ell_1$ norm at $x$

For any vector $u$

$$\|x + u\|_1 \geq \|x\|_1 + q^T u$$
Dual certificate

For any $x + h$ such that $Ah = 0$

$$\|x + h\|_1 \geq \|x\|_1 + q^T h$$

$q$ is a subgradient
Dual certificate

For any $x + h$ such that $Ah = 0$

\[ ||x + h||_1 \geq ||x||_1 + q^T h \]
\[ = ||x||_1 + v^T Ah \]

($q$ is a subgradient)

($q = A^T v$)
Dual certificate

For any $x + h$ such that $Ah = 0$

\[
\|x + h\|_1 \geq \|x\|_1 + q^T h
\]

\[
= \|x\|_1 + v^T Ah
\]

\[
= \|x\|_1
\]

(q is a subgradient)

(q = $A^T \nu$)
For any $x + h$ such that $Ah = 0$

$$
\|x + h\|_1 \geq \|x\|_1 + q^T h \quad (q \text{ is a subgradient})
$$

$$
= \|x\|_1 + v^T Ah
$$

$$
= \|x\|_1 \quad (q = A^T v)
$$

If $A_T$ (where $T$ is the support of $x$) is injective, $x$ is the unique solution
We need to \textbf{interpolate the sign} of an arbitrary well-separated signal with vectors in the \textit{row space} of \( A \).
Strategy

We need to **interpolate the sign** of an arbitrary well-separated signal with vectors in the **row space** of $A$

Correlation function $A^T A_i$ is in the row space!

$(A_i = i\text{th col of } A)$
We need to **interpolate the sign** of an arbitrary well-separated signal with vectors in the **row space** of $A$

Correlation function $A^T A_i$ is in the row space! ($A_i = i$th col of $A$)

**Proof of exact recovery:**

- Use correlations to interpolate
- Show that if separation is sufficient this yields valid certificate
Dual certificate construction
Dual certificate construction
Dual certificate construction
Dual certificate construction
Guarantees for SNL problems with decaying correlation

Theorem [Bernstein, Liu, Papadaniil, F. 2019]

In 1D, for any SNL problem with decaying correlation, $\ell_1$-norm minimization achieves exact recovery as long as the true parameters are sufficiently separated with respect to the correlation

- Result proved for continuous version of $\ell_1$ norm
- Additional condition: Decay of derivatives of correlation function
- Proof technique generalizes to higher dimensions
Dual certificate in higher dimensions
Robustness to noise / outliers

Variations of dual certificates establish robustness at small noise levels (Candès, F. 2013), (F. 2013), (Bernstein, F. 2017)

Exact recovery with constant number of outliers (up to log factors) (F., Tang, Wang, Zheng 2017), (Bernstein, F. 2017)

Open questions: Analysis of higher-noise levels and discretization error, robustness for positive amplitudes
Sparse recovery beyond compressed sensing: Separable nonlinear inverse problems. B. Bernstein, S. Liu, C. Papadaniil, C. Fernandez-Granda
Application to magnetic-resonance fingerprinting

Supported by a Moore-Sloan Data Science Environment seed grant and NIH R21 EB027241

Joint work with Jakob Assländer, Brett Bernstein, Martijn Cloos, Quentin Duchemin, Cem Gutelkin, Vlad Kobzar, Florian Knoll, Sylvain Lannuzel, Riccardo Lattanzi, and Sunli Tang
Quantitative MRI via fingerprinting

Radio-frequency pulses are designed to produce irregular magnetization signals (fingerprints) encoding relaxation parameters

![Graph showing time (s) vs. φ(T₁) with different T₁ values (120, 300, 720, 3200 ms)]
Multicompartiment magnetic resonance fingerprinting

- Assumption in MRF: One tissue per voxel
- Problematic at tissue boundaries
- Ignores sub-voxel structure
Additive model: Separable nonlinear inverse problem

\[ \phi(\theta_1) \]
\[ \phi(\theta_2) \]

Data
Correlation structure

![Graph showing correlation coefficients against $T_1$ (s)]
Multicompartment MRF via $\ell_1$-norm regularization

- Fast-thresholding methods don’t work
- We use an efficient interior-point solver
- Solving sequence of reweighted problems improves the solution

Drawback: Very slow
Validation with phantom
Validation with phantom

solutions A and C

solutions A and D

solutions B and D

1st
2nd
SC
GS
Current research

Goal: Fast multicompartment MRF for non-additive model

- Measurement design via ODE-constrained optimization
- Parameter estimation using a feedforward deep neural network trained on simulated data
For more information


**Multicompartment magnetic resonance fingerprinting.** S. Tang, C. Fernandez-Granda, S. Lannuzel, B. Bernstein, R. Lattanzi, M. Cloos, F. Knoll and J. Asslaender. Inverse Problems 34 (9) 4005. 2018

**Hybrid-State Free Precession for Measuring Magnetic Resonance Relaxation Times in the Presence of B0 Inhomogeneities.** V. Kobzar, C. Fernandez-Granda, J. Asslaender. ISMRM 2019
Separable nonlinear inverse problems

Machine learning for inverse problems
Data-driven estimation of sinusoid frequencies

Joint work with Brett Bernstein, Gautier Izacard, and Sreyas Mohan
Spectral super-resolution

Infinite samples

$N = 40$

$N = 20$
Traditional methodology

- Linear estimation (periodogram)
- Parametric methods based on eigendecomposition of sample covariance matrix (MUSIC, ESPRIT, matrix pencil)
- Sparsity-based methods
Learning-based approach
Comparison to state of the art

Blind denoising via convolutional neural networks

Joint work with Zahra Kadkhodaie, Sreyas Mohan, and Eero Simoncelli
**Goal:** Estimate image $x$ from data $y := x + z$ ($z$ is noise)

Feedforward convolutional neural networks are the state of the art
Image denoising via deep learning

**Goal:** Estimate image $x$ from data $y := x + z$ ($z$ is noise)

Feedforward convolutional neural networks are the state of the art

*Interesting phenomenon:* Removing additive constants in architecture provides generalization across noise levels

$$f(y) = W_L R(\ldots W_2 R(W_1 y + b_1) + b_2 \ldots) + b_L$$
Image denoising via deep learning

Goal: Estimate image $x$ from data $y := x + z$ ($z$ is noise)

Feedforward convolutional neural networks are the state of the art

*Interesting phenomenon:* Removing additive constants in architecture provides generalization across noise levels

$$f(y) = W_L R(\ldots W_2 R(W_1 y + b_1) + b_2 \ldots) + b_L$$
Generalization across noise levels

Training data (low noise)  Test image (high noise)  CNN  BF-CNN
Bias-free CNN is locally linear

\[ f(y) = W_L R W_{L-1} \ldots R W_1 y = A_y y \]

We can use linear-algebraic tools to visualize what is going on!
Rows interpreted as filters

Estimate at pixel $i$:

$$f_{BF}(y)(i) = (A_y y)(i) = \langle \text{ith row of } A_y, y \rangle$$
Low noise

Noisy image

Denoised

Pixel 1

Pixel 2

Pixel 3
Medium noise

Noisy image

Denoised

Pixel 1

Pixel 2

Pixel 3
High noise

Noisy image

Denoised

Pixel 1

Pixel 2

Pixel 3
Bias-free CNNs implement adaptive filters

Estimate at pixel $i$:

$$f_{BF}(y)(i) = (A_y y)(i) = \langle i\text{th row of } A_y, y \rangle$$

Rows can be interpreted as filters *adapted to image structure and noise*

Connection to classical Wiener denoising and nonlinear filtering
SVD analysis

\[ A_y = U S V^T \]

Empirical observations:

- Matrix is approximately symmetric \( U \approx V \)
- Matrix is approximately low-rank
Singular vectors computed from noisy image

Clean image

Large singular values

Small singular values
Bias-free CNNs enforce union-of-subspaces prior

\[ A_y \approx U S U^T \]

Low-dimensional subspace captures image features

BF-CNN implements union-of-subspaces prior

Connection to sparsity-based denoising
Robust and interpretable blind image denoising via bias-free convolutional neural networks
S. Mohan, Z. Kadkhodaie, E. Simoncelli, C. Fernandez-Granda
Conclusion

Analysis of $\ell_1$-norm minimization based on correlation decay and signal separation (as opposed to sparsity and incoherence)

Impressive empirical performance of machine-learning methods

Local linear-algebraic analysis reveals connections to existing techniques

**Challenge:** Develop mathematical understanding of ML methods!