A Second-Order Method for Sparse Signal Reconstruction in Compressed Sensing

Ioannis Dassios Kimon Fountoulakis Jacek Gondzio

University of Edinburgh

SIAM Conference on Optimization 2014

Outline

Sparse signal reconstruction problems in Compressed Sensing

- ℓ_1-analysis
- Total variation (special case of ℓ_1-analysis)

The method

- Primal-dual Newton Conjugate Gradients (modified) by **Chan, Golub, Mulet**. In *"A nonlinear primal-dual method for total variation-based image restoration."* *SIAM. J. Sci. Comput.* 20 (6) 1999 pp. 1964-1977.

Contribution

- Global and local convergence theory of pdNCG
- Preconditioning
- Potential of second-order methods for large-scale CS.

ℓ_1-analysis

$$\text{minimize } f_c(x) := c\|W^*x\|_1 + \frac{1}{2}\|Ax - b\|_2^2$$

- $x \in \mathbb{R}^n$
- $c \in \mathbb{R}_+$
- $W \in E^{n \times l}$, where $E = \mathbb{R}$ or $\mathbb{C}$
- $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $m \ll n$

Assumptions

- There is a unique solution x_c and W^*x_c has $q \ll l$ nonzeros.
- $\mathcal{N} := \{\text{indices of zeros in } W^*x_c\}$ and $W^*_{\mathcal{N}} := \{\text{rows of } W^* \text{ with indices in } \mathcal{N}\}$
  - $\text{Ker}(W^*_{\mathcal{N}}) \cap \text{Ker}(A) = \{0\}$.

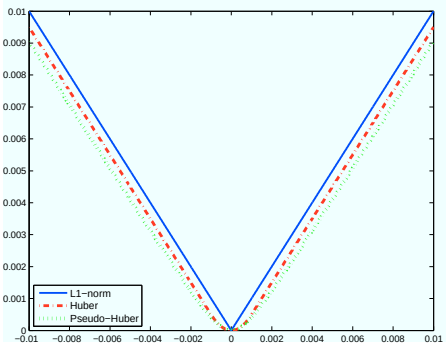
Difficulties

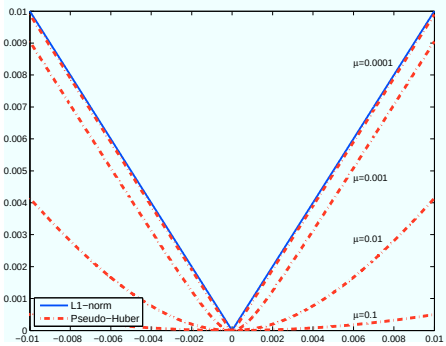
- Nonsmooth and nonseparable regulariser
- Frequently, large scale problems

Smoothing

- Huber: First-order differentiable
- **Pseudo-Huber**: *Second-order differentiable*

$$\psi_\mu(W^*x) = \sum_{i=1}^{l} \left(\sqrt{\mu^2 + (\overline{W}_i^* x)(W_i^* x)} - \mu \right)$$





Perturbation Analysis

ℓ_1-**analysis solution:** $x_c := \arg\min_x f_c(x)$

perturbed solution: $x_{c,\mu} := \arg\min_x f_c^\mu(x)$, where $f_c^\mu(x) = c\psi_\mu(W^*x) + \frac{1}{2}\|Ax - b\|_2^2$

Assumption: $\mathrm{Ker}(W_{\mathcal{N}}^*) \cap \mathrm{Ker}(A) = \{0\}$,
where $\mathcal{N}$ is the set of indices of the zero components in W^*x_c.

Theorem (Brief description). *There exists a $\tilde{\mu}$ such that $\forall \mu \leq \tilde{\mu}$ the difference of the two solutions is*

$$\|x_{c,\mu} - x_c\|_2 = \mathcal{O}(\mu^{1/2}) \quad \forall c, \mu.$$

I. Dassios, K. Fountoulakis & J. Gondzio,
"A Second-order Method for Compressed Sensing Problems with Coherent and Redundant Dictionaries." Technical Report ERGO-14-007

Shortcomings of smoothing in Newton

First-order optimality conditions

$$\nabla f_c^{\mu}(x) = c\underbrace{\frac{1}{2}(W_{re}DW_{re}^{\mathsf{T}} + W_{im}DW_{im}^{\mathsf{T}})x}_{\nabla\psi_{\mu}(W^*x)} + A^{\mathsf{T}}(Ax - b) = 0,$$

where $W_{re} := Re(W)$, $W_{im} := Im(W)$, $D := diag(D_1, D_2, \cdots, D_l)$ with

$$D_i := (\mu^2 + |W_i^* x|^2)^{-\frac{1}{2}} \quad \forall i = 1, 2, \cdots, l.$$

- For small μ, the gradient $\nabla\psi(W^*x)$ is highly nonlinear!
- Linearisation of $\nabla\psi(W^*x)$ by Newton method is inaccurate.
- Numerical instability and
- substantial shrinking of the region of convergence of Newton method.

A better linearisation

Set $g_{re} = DW_{re}^{\mathsf{T}} x$ and $g_{im} = DW_{im}^{\mathsf{T}} x$ in

$$\nabla f_c^{\mu}(x) = \frac{c}{2}(W_{re} DW_{re}^{\mathsf{T}} + W_{im} DW_{im}^{\mathsf{T}})x + A^{\mathsf{T}}(Ax - b) = 0,$$

and linearise

$$c(W_{re} g_{re} + W_{im} g_{im}) + A^{\mathsf{T}}(Ax - b) = 0,$$
$$D^{-1} g_{re} = W_{re}^{\mathsf{T}} x, \quad D^{-1} g_{im} = W_{im}^{\mathsf{T}} x.$$

These are perturbed optimality conditions of the primal-dual problem

$$\begin{array}{ll} \min\limits_{x \in \mathbb{R}^n} \max\limits_{g_{re}, g_{im} \in \mathbb{R}^l} & c\,(g_{re}^{\mathsf{T}} W_{re} x + g_{im}^{\mathsf{T}} W_{im} x) + \frac{1}{2}\|Ax - b\|_2^2 \\ \text{subject to:} & \|g_{re} + \sqrt{-1}\, g_{im}\|_\infty \leq 1. \end{array}$$

It has been observed by Chan, Golub, Mulet in
SIAM. J. Sci. Comput. 20 (6) 1999 pp. 1964-1977,

a dramatic improvement in the robustness of Newton method, even for small μ.

Determining the primal-dual directions

Linearisation of the new optimality conditions reduces to

$$B(x, g_{re}, g_{im})\Delta x = -\nabla f_c^{\mu}(x) \quad \text{where} \quad B := c\tilde{B}(x, g_{re}, g_{im}) + A^{\mathsf{T}}A. \tag{1}$$

The calculation of the dual directions Δg_{re} and Δg_{im} is inexpensive.

Three issues

- $\tilde{B}$ is not always symmetric.
- $\tilde{B}$ is positive definite if $\|g_{re} + \sqrt{-1}g_{im}\|_\infty \leq 1$.
- Solution of (1) is expensive.

Solution

- In (1), replace nonsymmetric B with symmetric $\bar{B} := c/2(\tilde{B} + \tilde{B}^{\mathsf{T}}) + A^{\mathsf{T}}A$.
- Maintain $\|g_{re} + \sqrt{-1}g_{im}\|_\infty \leq 1$, which implies $\bar{B} \succ 0$.
- Solve the linear system (1) approximately using PCG until

$$\|\bar{B}\Delta x + \nabla f_c^{\mu}(x)\|_2 \leq \eta \|\nabla f_c^{\mu}(x)\|_2, \quad \eta \in (0, 1).$$

Primal-dual Newton Conjugate Gradient (pdNCG)

1: **Input:** x^0, g_{re}^0 and g_{im}^0, where $\|g_{re}^0 + \sqrt{-1}g_{im}^0\|_\infty \leq 1$.

2: **Loop:** For $k = 1, 2, ...,$ until termination criteria are met.

3: Calculate primal-dual directions Δx^k, Δg_{re}^k and Δg_{im}^k approximately with PCG

4: Set $\tilde{g}_{re}^{k+1} := g_{re}^k + \Delta g_{re}^k$, $\tilde{g}_{im}^{k+1} := g_{im}^k + \Delta g_{im}^k$ and calculate

$$\bar{g}^{k+1} := P_{\|\cdot\|_\infty \leq 1}(\tilde{g}_{re}^{k+1} + \sqrt{-1}\tilde{g}_{im}^{k+1}),$$

where $P_{\|\cdot\|_\infty \leq 1}(\cdot)$ is the orthogonal projection on the ℓ_∞ ball.
Then set $g_{re}^{k+1} := Re\bar{g}^{k+1}$ and $g_{im}^{k+1} := Im\bar{g}^{k+1}$.

5: Perform backtracking line search on the primal direction to obtain a step-size α, which guarantees decrease of the primal objective function.

6: Set $x^{k+1} := x^k + \alpha\Delta x^k$.

Convergence analysis of pdNCG

Theorem (Primal convergence). *Let* $\{x^k\}_{k=0}^{\infty}$ *be a sequence generated by pdNCG. Then the sequence* $\{x^k\}_{k=0}^{\infty}$ *converges to the primal perturbed solution* $x_{c,\mu}$.

Theorem (Dual convergence). *The sequences of dual variables generated by pdNCG satisfy* $\{g_{re}^k\}_{k=0}^{\infty} \to D\mathrm{Re}W^{\intercal}x_{c,\mu}$, $\{g_{im}^k\}_{k=0}^{\infty} \to D\mathrm{Im}W^{\intercal}x_{c,\mu}$, *where* D *is measured at* $x_{c,\mu}$.

Lemma (Convergence of approximate Hessian). *Let the sequences* $\{x^k\}_{k=0}^{\infty}$, $\{g_{re}^k\}_{k=0}^{\infty}$ *and* $\{g_{im}^k\}_{k=0}^{\infty}$ *be generated by pdNCG. Then* $\bar{B}(x^k, g_{re}^k, g_{im}^k) \to \nabla^2 f_c^{\mu}(x_{c,\mu})$.

Theorem (Rate of convergence). *If* η^k *satisfies* $\lim_{k\to\infty} \eta^k = 0$, *then pdNCG converges superlinearly.*

I. Dassios, K. Fountoulakis & J. Gondzio,
"A Second-order Method for Compressed Sensing Problems with Coherent and Redundant Dictionaries." Technical Report ERGO-14-007

Preconditioner: intuition

Claim at the limit $k \to \infty$: $\|\bar{B}\|_2 = \|c/2(\tilde{B} + \tilde{B}^\intercal) + A^\intercal A\|_2 \approx \|c/2(\tilde{B} + \tilde{B}^\intercal)\|_2$

Prior information for $\mathbf{W^* x_c}$

- $q \ll l$ components are non-zero
- the majority $l - q$ components are zero

Information for $\mathbf{W^* x_{c,\mu}}$

For sufficiently small μ & $\|x_{c,\mu} - x_c\| = \mathcal{O}(\mu^{1/2})$, we expect that

- q components of $W^* x_{c,\mu}$ are non-zero
- the majority $l - q$ are $\mathcal{O}(\mu^{1/2})$

Information for $c/2(\tilde{B} + \tilde{B}^\intercal)$

- many eigenvalues are $\mathcal{O}(\frac{c}{\mu})$ (large!!)

Preconditioner and spectral properties

Approximate: $\bar{B} := c/2(\tilde{B} + \tilde{B}^\mathsf{T}) + A^\mathsf{T}A$,

with: $N := c/2(\tilde{B} + \tilde{B}^\mathsf{T}) + \rho I_m$, where $\rho \in [\delta_q, 1/2]$ and $\delta_q < 1/2$.

Two additional assumptions

- **Rows** of A are nearly orthogonal, i.e. $\|AA^\mathsf{T} - I_n\|_2 \leq \delta$, where δ is small.
- There exists $\delta_q < 1/2$ such that **Restricted Isometry Property (W-RIP)** holds:

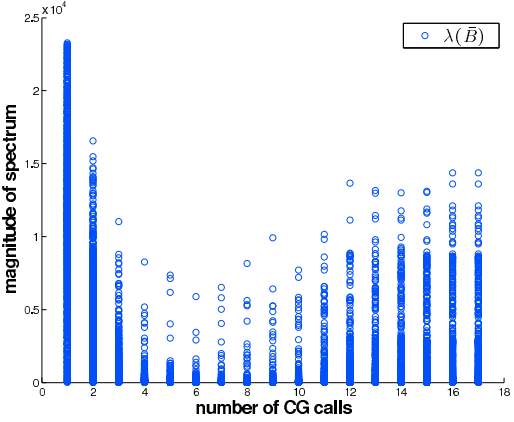
$$(1 - \delta_q)\|Wz\|_2^2 \leq \|AWz\|_2^2 \leq (1 + \delta_q)\|Wz\|_2^2,$$

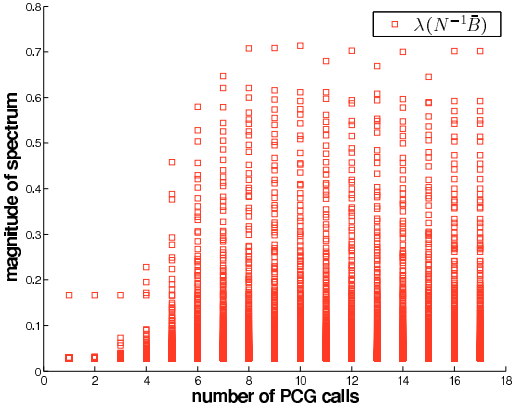
for all at most q-sparse $z \in E^l$. *In column spaces defined by any at most q columns of W matrix $A^\mathsf{T}A$* behaves like a scaled identity.

Theorem (Brief description). *Let $\lambda \in spec(N^{-1}\bar{B})$, then close to the solution $x_{c,\mu}$ the following holds*

- $|\lambda - 1| \leq \frac{1}{2}(\chi + 1 + (5\chi^2 - 2\chi + 1)^{\frac{1}{2}})\mathcal{O}(\mu)$, *where* $\chi := 1 + \delta - \rho$.

Spectral properties in practise ($\mu = 1.0e\text{-}5$)





pdNCG with preconditioning required much less CPU time

Solving systems with N

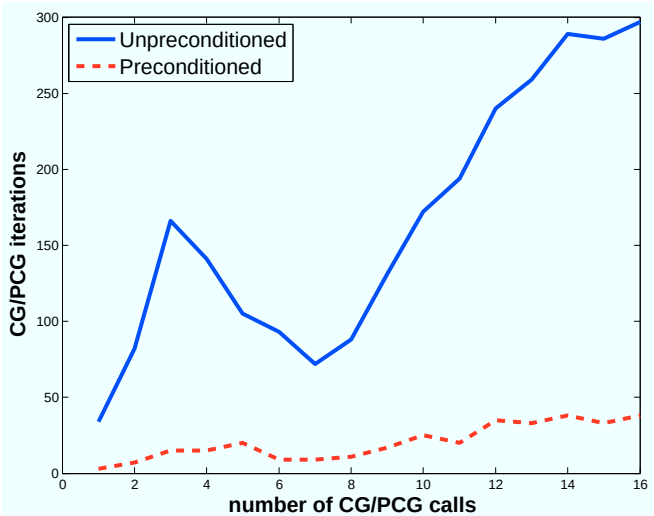
- **For Total-Variation:**
  N is a **5-diagonal** matrix (inexpensive to store and solve systems with)

- **For ℓ_1-analysis where the matrix W is arbitrary:**
  N does not have a structure (expensive)

  Solution by: **Vogel, Oman**, in *"Fast, Robust Total Variation Based Reconstruction of Noisy, Blurred Images." Image Processing, IEEE Transactions on* 7 (6) 1998 pp. 813-824.

  Briefly, if matrix-vector products with matrix W (hence with N) are inexpensive, then we can solve systems with matrix N inexactly using CG.

Efficiency of PCG ($\mu = 1.0e\text{-}5$)



Again, pdNCG with preconditioning required much less CPU time

Compared Solvers & Setting

We compare pdNCG with the first-order method TFOCS

- *TFOCS: Templates for First-Order Conic Solvers*
  by S. R. Becker, E. J. Candés and M. C. Grant
  - Algorithm: Auslender and Teboulle's single-projection method + smoothing of the ℓ_1-norm.

- **We make sure that problems are not over solved.**
- **We tune TFOCS according to comments of its authors.**
- **Experiments can be repeated by downloading the software from**
  `http://www.maths.ed.ac.uk/ERGO/pdNCG/`

Total-Variation

Info

- 256×256 phantom image
- A is a partial 2D discrete cosine transform (DCT)
- Measurements $m = 0.25n$
- SNR 10 dB
- **pdNCG**: time=16 sec., rel. err.=5.20e-1
- **TFOCS**: time=60 sec., rel. err.=5.20e-1

pdNCG

TFOCS

Single pixel camera problem set (http://dsp.rice.edu/cscamera)

Comparison on reconstruction of images which have been sampled by a single-pixel camera. There are five 64×64 images in total. The problems are reconstructed by Total-Variation. Matrix A is a partial Walsh basis which takes $0/1$ values with $m \approx 0.4n$ rows.

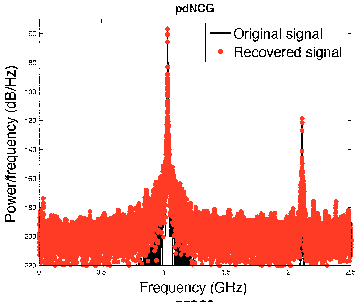
Results

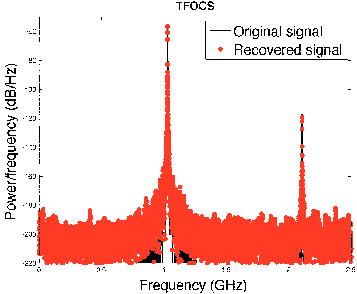
- pdNCG was faster on $4/5$ problems. On these problems on average pdNCG was 1.4 times faster.
- TFOCS was 1.3 times faster on the "R" image.

ℓ_1-Analysis (radio-frequency radar tones)

Info

- W is a $2^{15} \times 915456$ Gabor frame
- A is a 2616×2^{15} block diagonal matrix, with ± 1 for entries
- The small pulse has SNR 2.1e-2 dB, and the large pulse has SNR 60 dB
- **pdNCG**:
  time=420 sec., rel. err.=5.80e-4
- **TFOCS**: (converged after 150 iter.)
  time=615 sec., rel. err.=7.27e-4





Conclusion: **Careful exploitation of second-order information can speed up your method!**

Thank you!


I. Dassios, K. Fountoulakis, and J. Gondzio.
A second-order method for compressed sensing problems with coherent and redundant dictionaries.
Technical Report ERGO-14-007.

Continuation Framework

1: *Outer loop: For* $k = 0, 1, 2, \ldots, \vartheta$, *produce* $(c^k, \mu^k)_{k=0}^{\vartheta}$.
2: *Inner loop: Approximately solve the subproblem*

$$\text{minimize } f_{c^k}^{\mu^k}(x)$$

using pdNCG by initializing it with the solution of the previous subproblem.

Why continuation? Inexpensive Control of Spectrum

Spectrum (in practise): $\gamma I_n \preceq \bar{B} \preceq (\mathcal{O}(\frac{c}{\mu}) + \lambda_{max}(A^\top A))I_n$

Example without continuation: $c = 10^{-2}$, $\mu = 10^{-5}$

What happens in practise?

- During all stages of pdNCG $\kappa(\bar{B}) = \mathcal{O}(\frac{10^3}{\gamma})$

Example with continuation: $c^0 = \mu^0$ and $c = 10^{-2}$, $\mu = 10^{-5}$

What happens in practise?

- early stages $\kappa(\bar{B}) = \mathcal{O}(\frac{1}{\gamma})$
- late stages $\kappa(\bar{B}) = \mathcal{O}(\frac{10^3}{\gamma})$

Enable inexpensive preconditioning for small $\frac{c^k}{\mu^k}$.

Example of continuation on a TV problem