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

Ioannis Dassios    Kimon Fountoulakis    Jacek Gondzio

University of Edinburgh

20th Conference of the IFORS
Aim

- Robust solver with low per iteration computational cost
**Sparse signal reconstruction problems in Compressed Sensing**

- $\ell_1$-analysis
- Total variation (special case of $\ell_1$-analysis)

**The method**


**Contribution**

- Global and local convergence theory of pdNCG
- Preconditioning
- Potential of second-order methods for large-scale CS.
$\ell_1$-analysis

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 $\mathbb{C}$
- $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $m \ll n$
- Optimal $x_c$ has a sparse image through $W^*$; $W^*x_c$ is sparse

**Difficulties**

- Nonsmooth and nonseparable regulariser
- Frequently, large scale problems
- **Pseudo-Huber**: Second-order differentiable

\[
\psi_\mu(W^*x) = \sum_{i=1}^{l} \left( \sqrt{\mu^2 + (W_i^*x)(W_i^*x) - \mu} \right)
\]
Shortcomings of smoothing in Newton

First-order optimality conditions

\[ \nabla f_{c}^{\mu}(x) = c \nabla \psi_{\mu}(W^*x) + A^\top(Ax - b) = 0, \]

- \( \nabla \psi(W^*x) \) is highly nonlinear!
- Linearisation of \( \nabla \psi(W^*x) \) is inaccurate.
- the region of convergence of Newton method shrinks.
A better linearisation

\[ \nabla f^\mu_c(x) = c \frac{1}{2} (W_{re}DW_{re}^T + W_{im}DW_{im}^T)x + A^T(Ax - b) = 0, \]

\[ \nabla \psi_\mu(W^*x) \]

Set \( g_{re} = DW_{re}^Tx \) and \( g_{im} = DW_{im}^Tx \), and linearise blue instead of red.

\[ c(W_{re}g_{re} + W_{im}g_{im}) + A^T(Ax - b) = 0, \quad c(W_{re}g_{re} + W_{im}g_{im}) + A^T(Ax - b) = 0, \]

\[ D^{-1}g_{re} = W_{re}^Tx, \quad D^{-1}g_{im} = W_{im}^Tx. \]

These are perturbed optimality conditions of the primal-dual problem

\[ \min_{x \in \mathbb{R}^n} \max_{g_{re}, g_{im} \in \mathbb{R}^l} c \left( g_{re}W_{re}x + g_{im}W_{im}x \right) + \frac{1}{2} \|Ax - b\|_2^2 \]

subject to:

\[ \|g_{re} + \sqrt{-1}g_{im}\|_\infty \leq 1. \]

It has been observed by Chan, Golub, Mulet in


a dramatic improvement in the robustness of Newton method, even for small \( \mu \).
Determining the primal-dual directions

Linearisation of the new optimality conditions reduces to

\[ B(x, g_{re}, g_{im}) \Delta x = -\nabla f^\mu_c(x) \quad \text{where} \quad B := c\tilde{B}(x, g_{re}, g_{im}) + A^\top A. \] (1)

The calculation of the dual directions \( \Delta g_{re} \) and \( \Delta g_{im} \) 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 \( \bar{B} := c/2(\tilde{B} + \tilde{B}^\top) + A^\top A. \)
- Maintain \( \|g_{re} + \sqrt{-1}g_{im}\|_\infty \leq 1 \), which implies \( \bar{B} \succeq 0 \).
- Solve the linear system (1) approximately using PCG until

\[ \|\bar{B}\Delta x + \nabla f^\mu_c(x)\|_2 \leq \eta\|\nabla f^\mu_c(x)\|_2, \quad \eta \in (0, 1). \]
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 $k = 1, 2, \ldots$, until termination criteria are met.

3. Calculate primal-dual directions $\Delta x^k$, $\Delta g_{re}^k$ and $\Delta g_{im}^k$ 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

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

   where $P_{\|\cdot\|_{\infty} \leq 1}(\cdot)$ is the orthogonal projection on the $\ell_\infty$ ball.

   Then set $g_{re}^{k+1} := \text{Re}\tilde{g}^{k+1}$ and $g_{im}^{k+1} := \text{Im}\tilde{g}^{k+1}$.

5. Perform backtracking line search on the primal direction.

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 \( \{g^k_{re}\}_{k=0}^\infty \rightarrow \text{Re}(\nabla \psi_\mu(W^*x_{c,\mu})) \), \( \{g^k_{im}\}_{k=0}^\infty \rightarrow \text{Im}(\nabla \psi_\mu(W^*x_{c,\mu})) \).

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

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

Claim at the limit $k \to \infty$: 

\[ \| \bar{B} \|_2 = \| c/2(\tilde{B} + \tilde{B}^T) + A^T A \|_2 \approx \| c/2(\tilde{B} + \tilde{B}^T) \|_2 \]

Prior information for $W^* x_c$
- $q \ll l$ components are non-zero
- the majority $l - q$ components are zero

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

Information for $c/2(\tilde{B} + \tilde{B}^T)$
- many eigenvalues are $O(\frac{c}{\mu})$ (large!!)
Preconditioner and spectral properties

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

Assumptions

- Rows of \( A \) are nearly orthogonal, i.e. \( \|AA^\top - I_n\|_2 \leq \delta \), where \( \delta \) 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 \( A^\top A \) behaves like a scaled identity.

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

- \( |\lambda - 1| \leq \frac{1}{2}(\chi + 1 + (5\chi^2 - 2\chi + 1)\frac{1}{2})O(\mu), \) where \( \chi := 1 + \delta - \rho. \)
Spectral properties in practise ($\mu = 1.0e-5$)

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 $\ell_1$-analysis where the matrix $W$ is arbitrary:**
  $N$ does not have a structure *(expensive)*
  

  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 ($\mu = 1.0e-5$)

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 $\ell_1$-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
Comparison on reconstruction of images which have been sampled by a single-pixel camera. There are five $64 \times 64$ 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.
\( \ell_1 \)-Analysis (radio-frequency radar tones)

**Info**

- \( W \) is a \( 2^{15} \times 915456 \) Gabor frame

- \( A \) is a \( 2^{16} \times 2^{15} \) block diagonal matrix, with \( \pm 1 \) 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!

Continuation Framework

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

2. **Inner loop:** Approximately solve the subproblem

   $$\min_{c_k} f_{c_k}^{\mu_k}(x)$$

   using pdNCG by initializing it with the solution of the previous subproblem.
Why continuation? Inexpensive Control of Spectrum

\[ \gamma I_n \preceq \bar{B} \preceq (O(\frac{e}{\mu}) + \lambda_{\text{max}}(A^T A)) I_n \]

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

What happens in practise?

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

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

What happens in practise?

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

Enable inexpensive preconditioning for small \( \frac{c^k}{\mu^k} \).
Example of continuation on a TV problem