A Second-Order Method for Strongly Convex L1-regularisation Problems

Kimon Fountoulakis   Jacek Gondzio

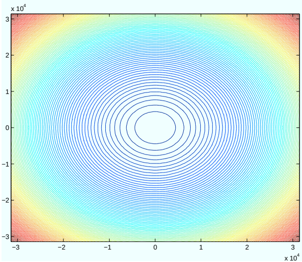
School of Mathematics
University of Edinburgh

4^{th} IMA Conference on Numerical Linear Algebra and Optimisation

September 4, 2014

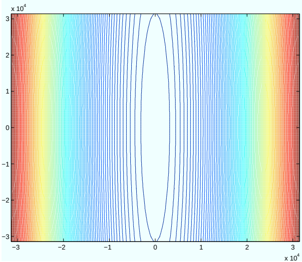
Problems of interest

Trivial



$$\nabla^2 f(x) = I_n$$

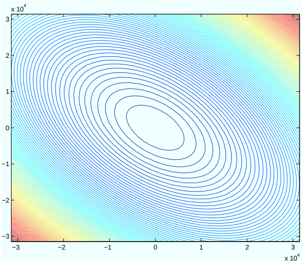
Inexpen. tractable



$$\nabla^2 f(x) = Q\Lambda Q^T$$

- Λ: large variations
- $Q_i \approx e_i$

Difficult



$$\nabla^2 f(x) = Q\Lambda Q^T$$

- Λ: large variations
- $Q_i \neq e_i$

Aim

- Robust solver with low per iteration computational cost

Outline

The method

- (pdNCG) 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
- Worst case iteration complexity
- Robust solver

Problem & Assumptions

$$\text{minimize } f_\tau(x) := \tau\|x\|_1 + \varphi(x)$$

- $x \in \mathbb{R}^n$, $\tau > 0$
- $\varphi(x) : \mathbb{R}^n \to \mathbb{R}$
- Optimal solution x_τ is sparse

A.1 $\varphi(x)$ is twice differentiable, and

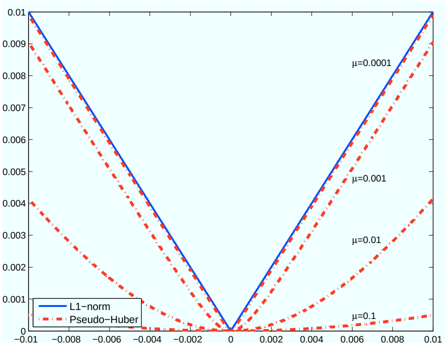
A.2 $\varphi(x)$ is strongly convex; at any x, $\lambda_n I_n \preceq \nabla^2\varphi(x) \preceq \lambda_1 I_n$

A.3 $\|\nabla^2\varphi(y) - \nabla^2\varphi(x)\| \leq L_\varphi\|y - x\|$

Addressing non-smoothness

Replace ℓ_1-norm with pseudo-Huber function

$$\psi_\mu(x) = \sum_{i=1}^{n} (\sqrt{\mu^2 + x_i^2} - \mu)$$



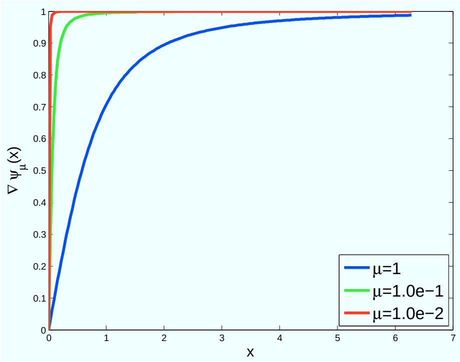
Shortcomings of smoothing in Newton

First-order optimality conditions

$$\nabla f_\tau^\mu(x) = \tau \underbrace{Dx}_{\nabla \psi_\mu(x)} + \nabla \phi(x) = 0,$$

where $D := diag(D_1, D_2, \cdots, D_n)$ with

$$D_i := (\mu^2 + x_i^2)^{-\frac{1}{2}} \quad \forall i = 1, 2, \cdots, n.$$



- $\nabla \psi(x)$ is highly nonlinear!
- Linearisation of $\nabla \psi(x)$ is inaccurate.
- the region of convergence of Newton method shrinks.

A better linearisation

Set $y := Dx$ in

$$\nabla f_\tau^\mu(x) = \tau Dx + \nabla \phi(x) = 0,$$

and linearise

$$\begin{aligned} \tau y + \nabla\phi(x) &= 0 \\ D^{-1} y &= x \end{aligned} \qquad \text{instead of} \qquad \begin{aligned} \tau y + \nabla\phi(x) &= 0 \\ y &= Dx \end{aligned}$$

These are perturbed optimality conditions of the saddle-point problem

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \max_{y\in\mathbb{R}^n} \quad & \tau\, y^\intercal x + \phi(x) \\ \text{subject to:} \quad & \|y\|_\infty \leq 1. \end{aligned}$$

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.

Primal-dual directions

Linearisation of the new optimality conditions reduces to

$$H(x,y)\Delta x = -\nabla f_\tau^\mu(x) \tag{1}$$

where

$$\begin{aligned} H &:= \tau \tilde{H}(x,y) + \nabla^2 \phi(x) \\ \tilde{H}(x,y) &= \tau D(I - D diag(x) diag(y)) \approx \nabla^2 \psi(x) \end{aligned}$$

Two issues

- $\tilde{H}$ is positive definite if $\|y\|_\infty \leq 1$.
- Solution of (1) is expensive.

Solution

- Maintain $\|y\|_\infty \leq 1$.
- Solve the linear system (1) inexactly using PCG

Primal-dual Newton Conjugate Gradient (pdNCG)

1: **Input:** x^0, y^0, where $\|y^0\|_\infty \leq 1$.

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

3: Calculate primal-dual directions Δx^k, Δy^k inexactly with PCG.

4: Set $\tilde{y}^{k+1} := y^k + \Delta y^k$ and calculate

$$y^{k+1} := P_{\|\cdot\|_\infty \leq 1}(\tilde{y}^{k+1}),$$

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

5: Perform backtracking line search on the primal direction.

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

Termination criterion

Definition of inexact Newton decrement: $\|\Delta x\|_x$

- Let Δx_i be the direction obtained by PCG at the i^{th} iteration,

$$\|\Delta x_i\|_x^2 := (\Delta x_i)^\intercal H(x,y)\Delta x_i = -(\Delta x_i)^\intercal \nabla f_\tau^\mu(x).$$

Intepretation

- Let $Q(x+\Delta x) := f_\tau^\mu(x) + \nabla f_\tau^\mu(x)^\intercal \Delta x + (\Delta x)^\intercal H(x,y)\Delta x$
- Then

$$1/2\|\Delta x_i\|_x^2 = f_\tau^\mu(x) - \min_{\Delta x \in \mathcal{K}_i} Q(x+\Delta x) \approx f_\tau^\mu(x) - f^* \ (\text{as } k \to \infty)$$

where

$$\mathcal{K}_i := \text{span}(\nabla f_\tau^\mu(x), H(x,y)\nabla f_\tau^\mu(x), \ldots, H(x,y)^{i-1}\nabla f_\tau^\mu(x))$$

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 $x_{\tau,\mu}$.*

Theorem (Dual convergence). *The sequences of dual variables generated by pdNCG satisfy $\{y^k\}_{k=0}^{\infty} \to \nabla \psi_\mu(x_{\tau,\mu})$.*

Lemma (Convergence of approximate Hessian). *Let the sequences $\{x^k\}_{k=0}^{\infty}$ and $\{y^k\}_{k=0}^{\infty}$ be generated by pdNCG. Then $H(x^k, y^k) \to \nabla^2 f_\tau^\mu(x_{\tau,\mu})$.*

Notation (before theory...)

- $f(x) := f_\tau^\mu(x)$
- L: Lipschitz constant of $\nabla^2 f(x)$
- $H(x, y)$ is uniformly bounded

$$\lambda_n I_n \preceq H(x, y) \preceq \lambda_1 I_n,$$

where $0 < \lambda_n \leq \lambda_{n-1} \leq \cdots \leq \lambda_1$.

pdNCG: global and local convergence behaviour

Minimum step-size and minimum decrease

$$\alpha \geq \mathcal{O}\left(\frac{\lambda_n}{\lambda_1}\right), \qquad f(x) - f(x(\alpha)) \geq \mathcal{O}\left(\frac{\lambda_n}{\lambda_1}\right) \|\Delta x\|_x^2$$

If $\|\Delta x\|_x \leq \varpi$, $0 < \varpi \leq c$, where

$$c = \mathcal{O}(\frac{\lambda_n^{\frac{3}{2}}}{L})$$

then $\alpha = 1$ and

$$\frac{1}{2}\frac{\mathcal{O}(\lambda_1^2, \lambda_n^{\frac{1}{2}}, L)}{\lambda_n^{\frac{3}{2}}}\|\Delta x^{k+1}\|_{x^{k+1}} \leq \Big(\frac{1}{2}\frac{\mathcal{O}(\lambda_1^2, \lambda_n^{\frac{1}{2}}, L)}{\lambda_n^{\frac{3}{2}}}\|\Delta x^k\|_{x^k}\Big)^2$$

Worst-case iteration complexity of pdNCG

$$\frac{f(x^0) - f(x_{\tau,\mu})}{\mathcal{O}(\frac{\lambda_n^2}{\lambda_1^2})} + \log_2 \log_2 \left(\frac{const.}{\epsilon}\right)$$

iterations to converge to a solution x^k of accuracy

$$f(x^k) - f(x_{\tau,\mu}) \leq \epsilon.$$

Standard Newton, see S. Boyd and L. Vandenberghe, *Convex Optimization*

$$\frac{f(x^0) - f(x_{\tau,\mu})}{\mathcal{O}(\frac{\lambda_n^5}{L^2\lambda_1^2})} + \log_2 \log_2 \frac{const.}{\epsilon}.$$

Sparse Least-Squares

$$\varphi(x) = \frac{1}{2}\|Ax - b\|^2$$

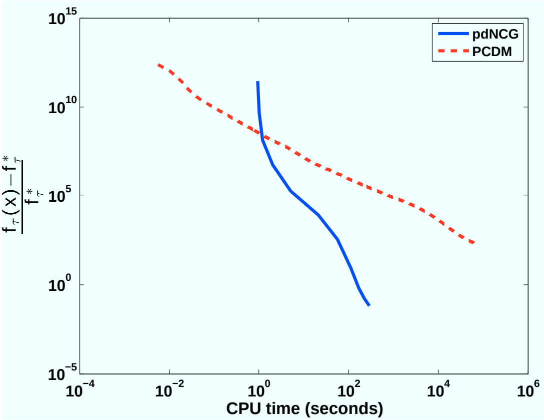
where $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$ with $m \geq n$.

- PCDM: Parallel Coordinate Descent Method
  - very efficient on problems that are well-scaled or poorly-scaled along co-ordinate axis
  - exploits multi-core systems
- A 40-core system was used.

Difficult small scale problem

Info

- $\tau = 1$
- $n = 4,096$
- $m = 1.01n$
- $cond(A^T A) = 9.0e{+}8$
- $nnz(A)/(mn) = 9.0e\text{-}3$



PCDM was terminated after 30 million iterations.

Large scale problem

Info

- $\tau = 1$
- $n = 2^{27} \approx 130$ m.
- $m = 1.1n$
- $nnz(A)/(mn) = 1.0e\text{-}8$

	pdNCG	**PCDM**
CPU sec.	$2,550$	$22,300$
rel. err	$3.79e$-03	$2.47e+06$
$f_\tau(x)$	$5.20e+05$	$1.14e+13$

Difficult Machine Learning problems

$$\varphi(x) = \sum_{i=1}^{m} \log(1 + e^{-y_i x^\intercal w_i}),$$

where $x, w_i \in \mathbb{R}^n$ are the feature vectors and $b_i \in \{-1, +1\}$ are the corresponding labels.

Problem	m	n	$nnz(W)/(mn)$	τ
cod-rna	$59,535$	8	$1.00e$-00	$1.11e$+01
covtype	$581,012$	54	$2.20e$-01	$4.58e$-02

- PCDM

- newGLMNET: Newton-type; obtains a direction at step k by solving approximately subproblem

$$d_k := \arg\min_d \tau \|x_k + d\|_1 + \nabla\phi(x_k)^T d + \frac{1}{2} d^T \nabla^2 \phi(x_k) d$$

using a co-ordinate descent method.

Difficult Machine Learning problems

Problem	**PCDM**		**newGLMNET**	
	$f_\tau(x)$	CPU sec.	$f_\tau(x)$	CPU sec.
cod-rna	$2.16e{+}05$	103	$2.16e{+}05$	0.7
covtype	$7.35e{+}05$	1530	$7.31e{+}05$	51

Problem	**pdNCG**	
	$f_\tau(x)$	CPU sec.
cod-rna	$2.27e{+}05$	0.3
covtype	$7.20e{+}05$	5.4

Conclusion

- Complete analysis of pdNCG.
- Numerical results which show that pdNCG is robust and efficient on difficult examples.

Thank You!

Kimon Fountoulakis and Jacek Gondzio.
A second-order method for strongly convex ℓ_1-regularization problems.
Technical Report ERGO-14-005, 2014.

Software: `http://www.maths.ed.ac.uk/ERGO/pdNCG/`