A Second-Order Method for Strongly Convex ℓ_1-Regularization Problems

Kimon Fountoulakis Jacek Gondzio

School of Mathematics
University of Edinburgh

26^{th} European Conference On Operational Research

July 3, 2013

Problem & Assumptions

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

- $x \in \mathbb{R}^m$, $\tau > 0$
- $\varphi(x) : \mathbb{R}^m \to \mathbb{R}$

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

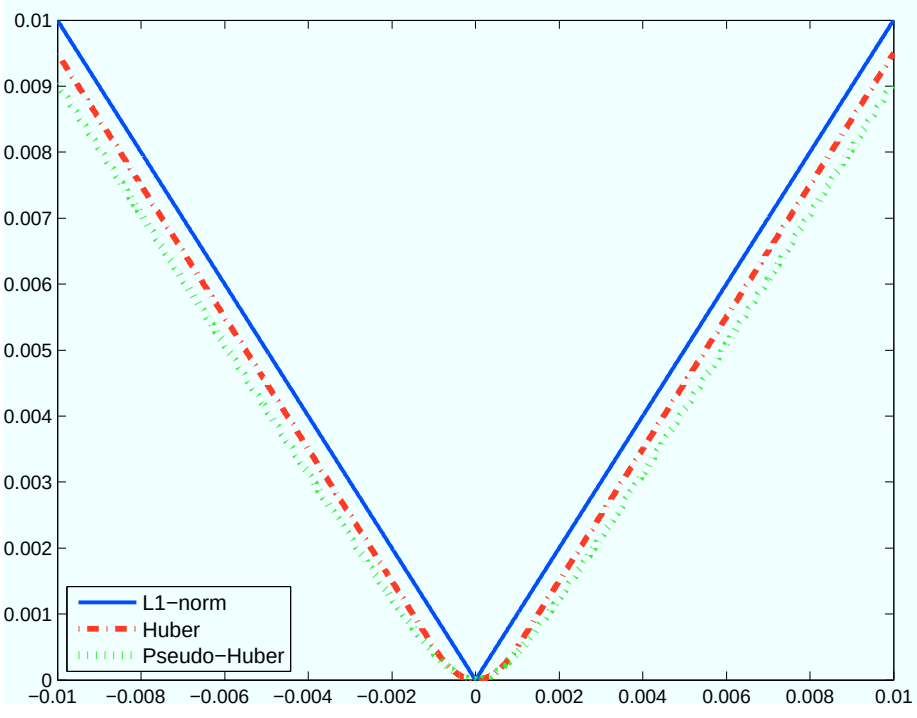
A.2 $\varphi(x)$ is strongly convex; at any x, $\lambda_m I_m \preceq \nabla^2\varphi(x) \preceq \lambda_1 I_m$

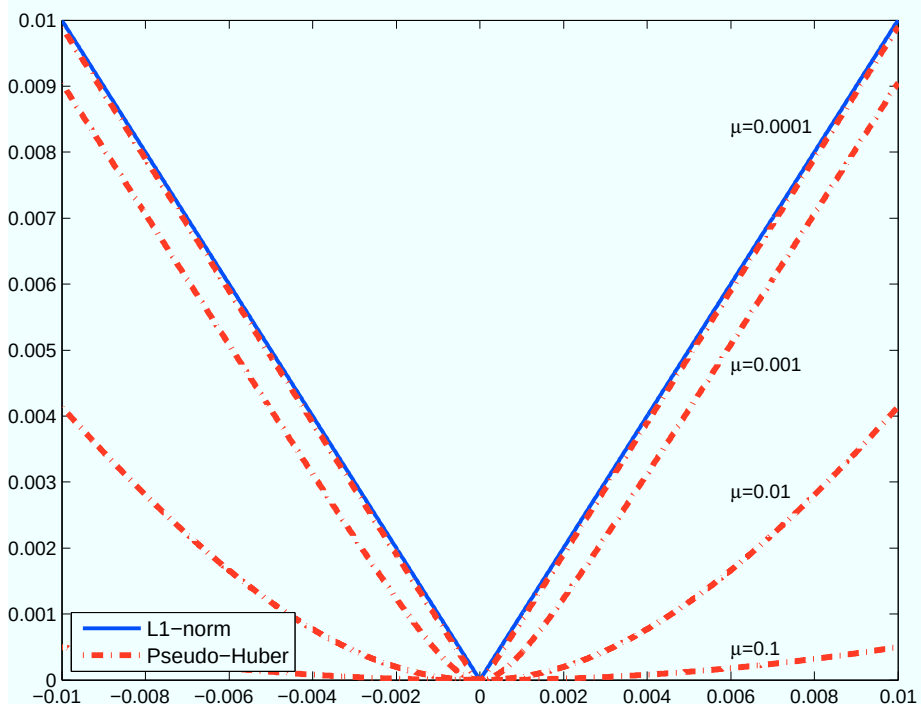
A.3 $\|\nabla^2\varphi(y) - \nabla^2\varphi(x)\| \leq L_\varphi\|y - x\|$

Addressing non-smoothness

Replace ℓ_1-norm with pseudo-Huber function

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





Doubly-Continuation N-CG

Replace $\min_{x \in \mathbb{R}^m} \; f_\tau(x) := \tau \|x\|_1 + \varphi(x)$

with $\min_{x \in \mathbb{R}^m} \; f_\tau^\mu(x) := \tau \psi_\mu(x) + \varphi(x)$

Algorithm dcNCG

1: **Outer loop:** For $l = 1, 2, \ldots, \vartheta$ continuation iterations, produce

- $\{\tau^l\} \to \bar{\tau}$ and $\{\mu^l\} \to \bar{\mu}$
- $\tau^{l+1} = \beta_1 \tau^l$ and $\mu^{l+1} = \beta_2 \mu^l$, $\beta_1, \beta_2 \in [0, 1)$

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

$$\text{minimize } f_{\tau^l}^{\mu^l}(x)$$

using N-CG with backtracking line-search

Warm-start and control of spectrum

Warm-start

- **minimize** $\mathbf{f}_\tau(\mathbf{x})$ has the zero solution for $\boldsymbol{\tau} \geq \|\boldsymbol{\nabla}\boldsymbol{\varphi}(\mathbf{0})\|_\infty$
- **minimize** $\mathbf{f}_\tau^\mu(\mathbf{x})$ has nearly the zero solution for $\boldsymbol{\tau} \gg \|\boldsymbol{\nabla}\boldsymbol{\varphi}(\mathbf{0})\|_\infty$ and $\boldsymbol{\mu} \approx \mathbf{1}$

Control of spectrum

- At any x, $\lambda_m I_m \preceq \nabla^2 f_\tau^\mu(x) \preceq \left(\frac{\tau}{\mu} + \lambda_1\right) I_m$
- $\nabla^2 f_\tau^\mu(x) = \tau \nabla^2 \psi_\mu(x) + \nabla^2 \varphi(x)$

$$\nabla^2 \psi_\mu(x) = diag\left(\frac{1}{\mu}\left[(1 + \frac{x_1^2}{\mu^2})^{-\frac{3}{2}}, \ldots, (1 + \frac{x_n^2}{\mu^2})^{-\frac{3}{2}}\right]\right)$$

N-CG with backtracking line-search

Solve approximately $\nabla^2 f_{\tau^l}^{\mu^l}(x^k) d^k = -\nabla f_{\tau^l}^{\mu^l}(x^k)$

Local norm: $\|u\|_{x,l} = \sqrt{u^T \nabla^2 f_{\tau^l}^{\mu^l}(x) u}$

Algorithm Backtracking line-search

1: **Input** Given τ^l, μ^l and an inexact Newton direction d for point x, calculated by CG and $c_2 \in (0, \frac{1}{2})$, $c_3 \in (0, 1)$

2: set $\alpha = 1$

3: **while** $f_{\tau^l}^{\mu^l}(x + \alpha d) > f_{\tau^l}^{\mu^l}(x) - c_2 \alpha \|d\|_{x,l}^2$ **do**

4: $\quad \alpha = c_3 \alpha$

5: **end while**

Notation N-CG

For the analysis of N-CG standalone

- $f(x) := f_\tau^\mu(x)$, smooth and strongly-convex
- $L := \frac{\tau}{\mu^2} + L_\varphi$
- $\tilde{\lambda}_1 := \frac{\tau}{\mu} + \lambda_1$

Local norm: $\|u\|_x = \sqrt{u^T \nabla^2 f_\tau^\mu(x) u}$

N-CG: global and local convergence rates

Minimum step-size $\bar{\alpha}$ and minimum decrease

$$\bar{\alpha} \geq c_3 \frac{\lambda_m}{\tilde{\lambda}_1}, \qquad f(x) - f(x(\bar{\alpha})) \geq c_2 c_3 \frac{\lambda_m}{\tilde{\lambda}_1} \|d\|_x^2$$

If $\|d^k\|_{x^k} \leq \varpi$, $0 < \varpi \leq c_5$, where

$$c_5 = \min \left\{ 3(1 - 2c_2) \frac{\lambda_m^{\frac{3}{2}}}{L}, \frac{\lambda_m^{\frac{3}{2}}}{16\tilde{\lambda}_1^2 + L} \right\},$$

then $\bar{\alpha} = 1$ and

$$\frac{1}{2} \frac{16\tilde{\lambda}_1^2 + L}{\lambda_m^{\frac{3}{2}}} \|d^{k+1}\|_{x^{k+1}} \leq \left(\frac{1}{2} \frac{16\tilde{\lambda}_1^2 + L}{\lambda_m^{\frac{3}{2}}} \|d^k\|_{x^k} \right)^2.$$

Worst-case iteration complexity of N-CG

N-CG with backtracking line-search requires at most

$$K_3 = c_6 \log\left(\frac{f(x^0) - f(x^*)}{c_7 \varpi^2}\right) + \log_2 \log_2\left(\frac{c_8}{\epsilon}\right)$$

iterations to converge to a solution x^k, $k > 0$, of accuracy

$$f(x^k) - f(x^*) \leq \epsilon,$$

where $c_6 = \mathcal{O}(\kappa^3)$.

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

$$\frac{f(x^0) - f(x^*)}{\gamma} + \log_2 \log_2 \frac{\epsilon_0}{\epsilon}$$

where $\gamma = c_2 c_3 \frac{\lambda_m}{\tilde{\lambda}_1} \|d_N\|_x^2$ and $\epsilon_0 = 2\frac{\lambda_m^3}{L^2}$.

dcNCG: which parameter terminates first? τ or μ?

A.4 τ converges with a faster rate, but μ converges first.

Is this assumption important? No! But it guarantees

$$f_{\tau^0}^{\mu^0}(z^0) > f_{\tau^1}^{\mu^1}(z^0) > f_{\tau^1}^{\mu^1}(z^1) > \ldots > f_{\bar{\tau}}^{\bar{\mu}}(\bar{z}),$$

where z^l is the minimizer of $f_{\tau^l}^{\mu^l}(x)$.

Assumption $A.4$ is only used to obtain a worst-case complexity result of dcNCG independent of the continuation index l.

Worst-case iteration complexity of dcNCG

If every l^{th} subproblem is solved to accuracy

$$\|d^k\|_{x^k,l} \leq \epsilon^l, \quad \text{where} \quad \epsilon^l \leq \tilde{c}_5$$

and

$$\tilde{c}_5 = \min\left\{3(1-2c_2)\frac{\lambda_m^{\frac{3}{2}}}{\left(\frac{\tau^0}{\mu^0\bar{\mu}}+L_\varphi\right)}, \frac{\lambda_m^{\frac{3}{2}}}{16\left(\frac{\tau^0}{\mu^0}+\lambda_1\right)^2+\left(\frac{\tau^0}{\mu^0\bar{\mu}}+L_\varphi\right)}\right\},$$

then dcNCG algorithm with backtracking line-search initialized at x^0 requires at most

$$K_5 = \underbrace{\left(\tilde{c}_6 \log\left(\frac{\tilde{\rho}}{\tilde{c}_7\tilde{\varpi}^2}\right)+\log_2\log_2\left(\frac{\tilde{c}_8}{\tilde{\epsilon}}\right)\right)}_{\text{Worst-case } K_3^l \quad \forall l} \quad \underbrace{\log_{\beta_1}\frac{\bar{\tau}}{\tau^0}}_{\text{Continuation iter.}} .$$

Sparse Least-Squares

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

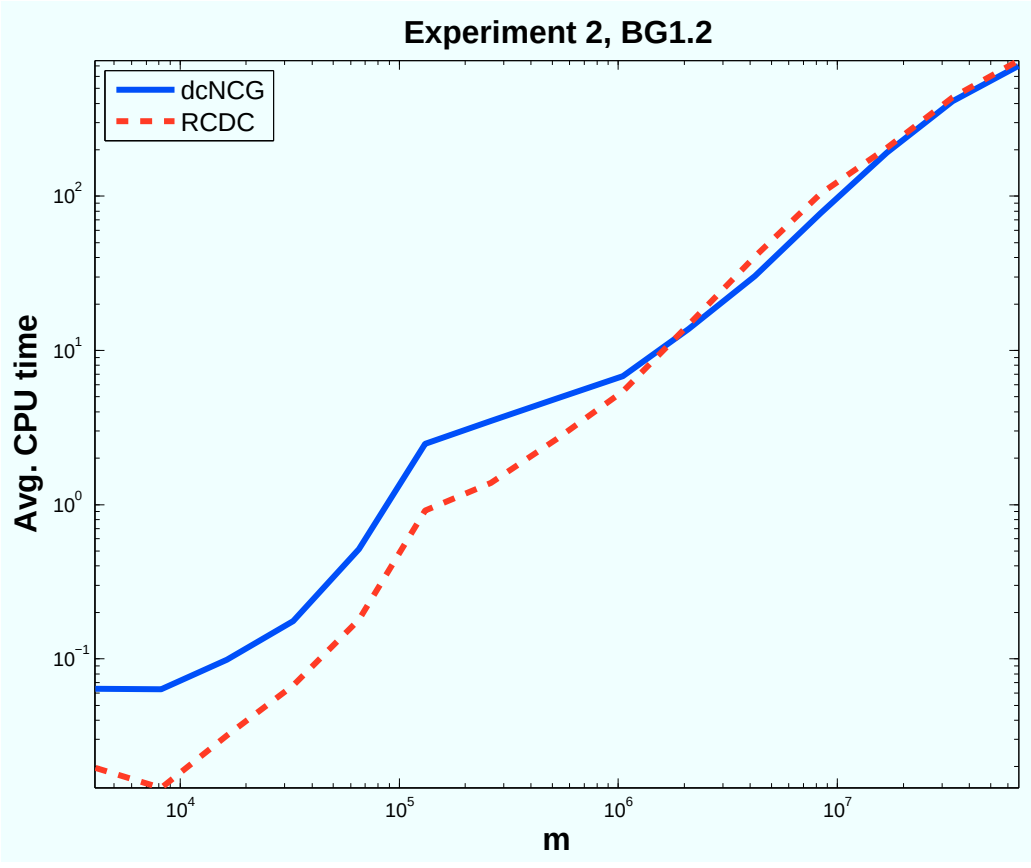
where $x \in \mathbb{R}^m$, $b \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times m}$ with $n \geq m$.

- RCDC: randomized parallel coordinate descent by P. Richtárik and M. Takáč in *Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function*.
  - exploits sparsity/separability of problem
  - exploits multi-core systems
- dcNCG implements Ax and $A^T y$ in parallel
- Experiments are run on a 24 core system
- BG1.2 have been first proposed by **Y. Nesterov**, *Gradient Methods for Minimizing Composite Objective Function*

Sparse Least-Squares: Increasing m

Info

- $\frac{\text{mass of } diag(A^T A)}{\text{mass of } A^T A} \approx 99.9\%$
- $n = 2m$
- 20 non-zero elements per column of A
- PCG with $P = diag(A^T A)$



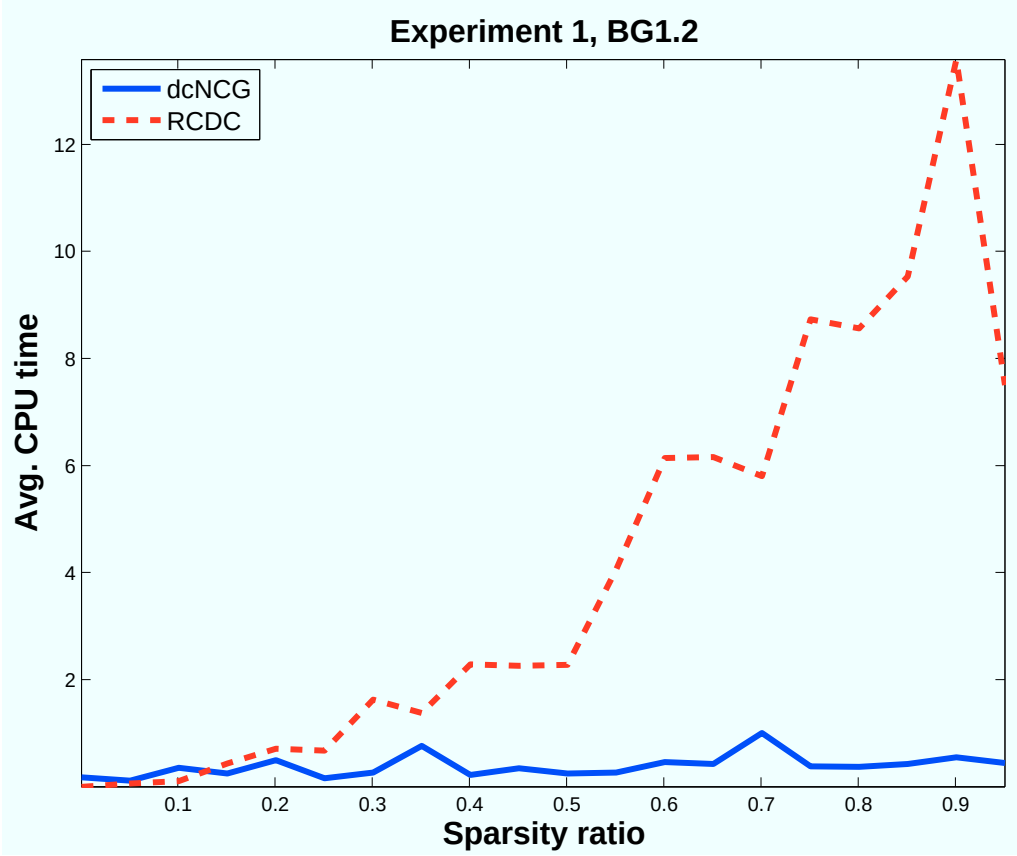
Why is RCDC so fast? Coordinate directions biased with second-order information

$$d_i := \arg\min_{p_i} \ \tau|x_i + p_i| + [\nabla\varphi(x)]_i p_i + \frac{\beta}{2} p_i [diag(A^T A)]_{ii} p_i$$

Sparse Least-Squares: Increasing density

Observations

- $\frac{\text{mass of } diag(A^\intercal A)}{\text{mass of } A^\intercal A} \geq 40\%$
- $cond(P^{-1}A^\intercal A) = 10^7$, where $P = diag(A^\intercal A)$
- dcNCG is not affected by the sparsity ratio or any properties of $diag(A^T A)$

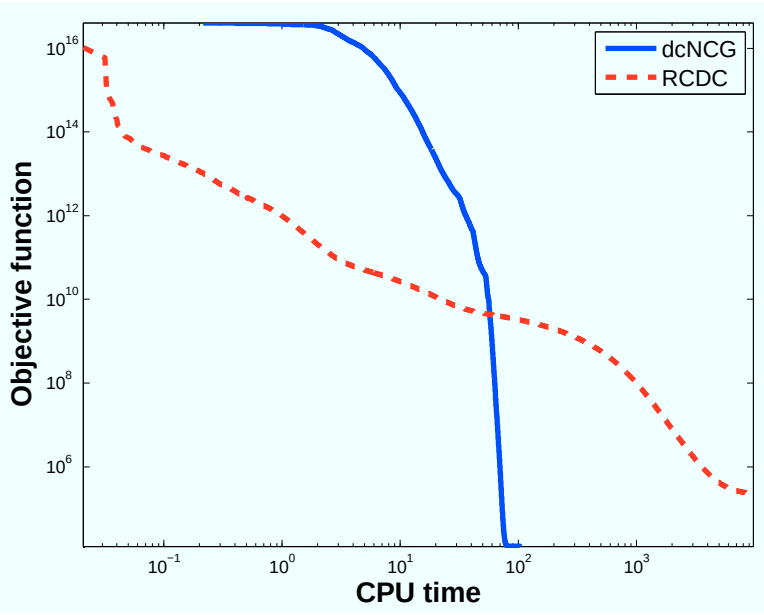


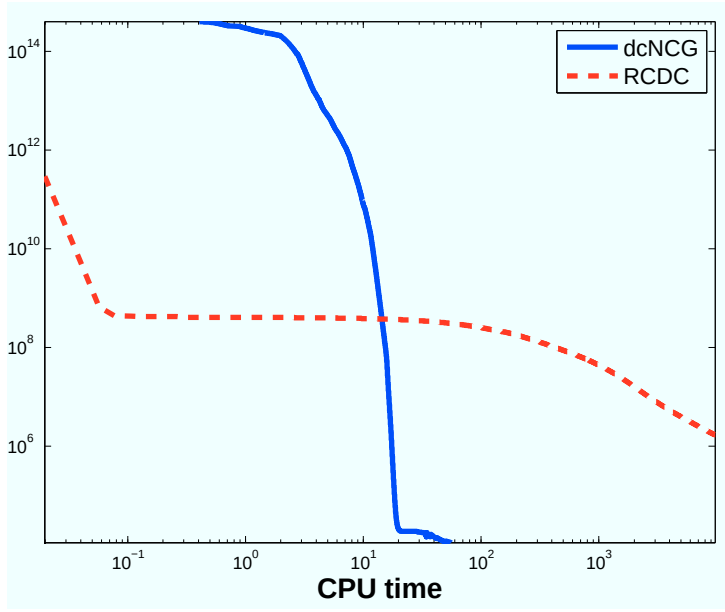
$m = 10^3$, $n = 2m$

$f(x) - f(x^*) = \mathcal{O}(10^{-6})$

sparsity ratio: # of non-zero elements per column of A

Sparse Least-Squares: Difficult cases





- The singular values of A are uniformly distributed in $(10^{-1}, 10^{6})$.
- Two clusters of singular values, 100 are 10^{5} and the rest are 10^{-1}.

x^* has 50 non-zero elements, randomly positioned with elements uniformly distributed in $(10^{-3}, 10^{3})$

$m = 2^{12}$ and $n = 2m$

RCDC was terminated after 10 million iterations $\approx 10k$ seconds

Logistic Regression

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

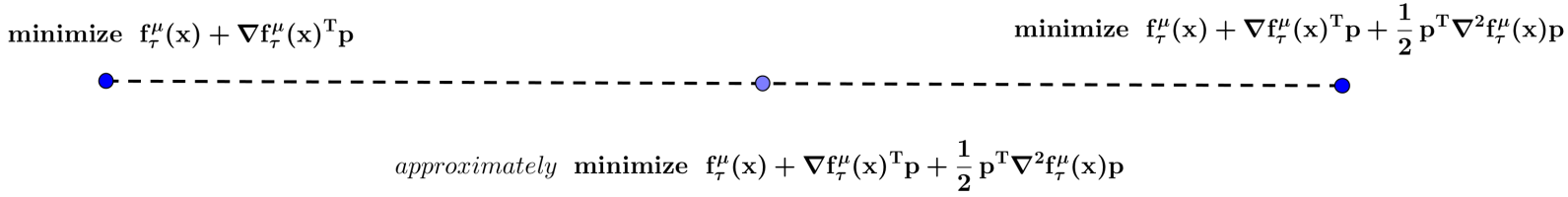
where $w, x_i \in \mathbb{R}^n$ are the feature vectors and $y_i \in \{-1, +1\}$ are the corresponding labels.

dcNCG was faster on 37 *out of* 45 *Logistic Regression problems of the LIBSVM package against CDN (Coordinated Descent Newton) of LIBLINEAR package.*

- Both solvers are run in serial mode and are C/C++ implementations
- Both solvers achieved same level of accuracy

Conclusions

- A worst-case iteration complexity of dcNCG is available.
- Extensive numerical results which show that dcNCG is efficient.
- dcNCG can balance between 1st and 2nd order methods.



Thank you!

Kimon Fountoulakis and Jacek Gondzio.
A second-order method for strongly convex ℓ_1-regularization problems.
Technical Report ERGO-13-011.

Termination of N-CG

Newton decrement

$$\|d_N\|_{x,I} = \inf\{c \mid |\nabla f(x)^T h| \leq c\|h\|_{x,I} \ \forall h \in \mathbb{R}^n\},$$

for $h = d_N$.

Standard Newton termination criterion $\|d_N\|_{x,I} \leq \epsilon$.

For d given by CG and $|\nabla f(x)^T d| = \|d\|_x^2$

$$\|d_N\|_{x,I} \leq \|d\|_{x,I} = \inf\{c \mid |\nabla f(x)^T d| \leq c\|d\|_{x,I}\}$$

N-CG can be terminated when $\|d\|_{x,I} \leq \epsilon$.

Yet another analysis of N-CG?

	Current	Our analysis	Newton
Global convergence	✓	✓	✓
Global linear conv. rate		✓	✓
Local quad. conv. region		✓	✓
Local quad. rate	✓	✓	✓
Worst-case complexity		✓	✓

- **S. Wright and J. Nocedal** *Numerical Optimization*, 2006
- **H. F. Walker et al.** *Globalization techniques for Newton-Krylov methods and applications to the fully-coupled solution of the Navier-Stokes equations*, 2006
- **S. G. Nash** *A survey of truncated-Newton methods*, 2000
- **P. N. Brown and Y. Saad** *Convergence theory of nonlinear Newton-Krylov algorithms*, 1994