Second-Order Methods for Signal Reconstruction, Big-Data and Machine Learning Problems

Ioannis Dassios Kimon Fountoulakis Jacek Gondzio

School of Mathematics
University of Edinburgh

ICCOPT 2013

July 30, 2013

Formulation & Assumptions

$$\text{minimize } f_c(x) := \sum_{i=1}^{p} c_i \|W_i x\|_1 + \varphi(x)$$

- $x \in \mathbb{R}^m$
- $c \in \mathbb{R}^p_+$
- $W_i : \mathbb{R}^m \to E^{l_i}$, where $E = \mathbb{R}$ or $\mathbb{C}$
- $\varphi(x) : \mathbb{R}^m \to \mathbb{R}$

A.1.1 $\varphi(x)$ is twice differentiable and convex, or

A.1.2 $\varphi(x)$ is strongly convex; at any x, $\lambda_m I_m \preceq \nabla^2 \varphi(x) \preceq \lambda_1 I_m$

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

Problems

Type	**p**	**W**	$\varphi(x)$
Least-Squares	1	Identity	$\|Ax - b\|_2^2$
Logistic-Regression	1	Identity	$\sum_{i=1}^{l} \log(1 + e^{-y_i x^\intercal w_i})$
ℓ_1-analysis	1	Arbitrary matrix $W : \mathbb{R}^m \to \mathbb{R}^l$	$\|Ax - b\|_2^2$
isotropic TV (iTV)	1	$D : \mathbb{R}^m \to \mathbb{C}^{m-1}$ (Tridiagonal matrix)	$\|Ax - b\|_2^2$
iTV & ℓ_1-analysis	2	$D : \mathbb{R}^m \to \mathbb{C}^{m-1}$ $W : \mathbb{R}^m \to \mathbb{R}^l$	$\|Ax - b\|_2^2$

Smoothing

Nesterov's smoothing for the ℓ_1-norm: Huber function (only first-order differentiable),

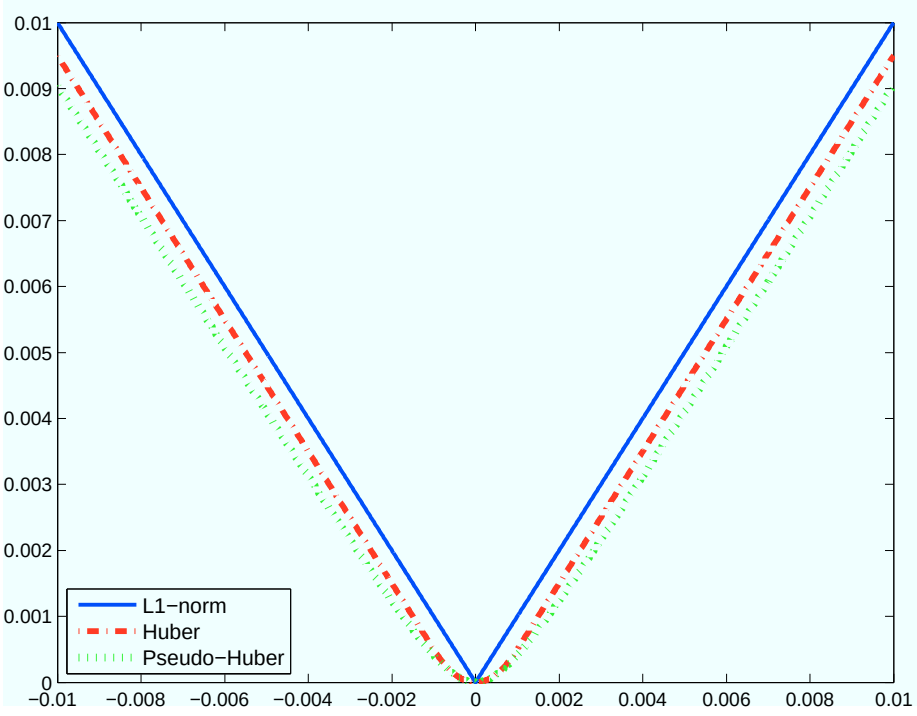
- used in NestA by S. R. Becker and J. Bobin and E. J. Candés,
- and in TFOCS by S. R. Becker and E. J. Candés and M. C. Grant, for the dual.

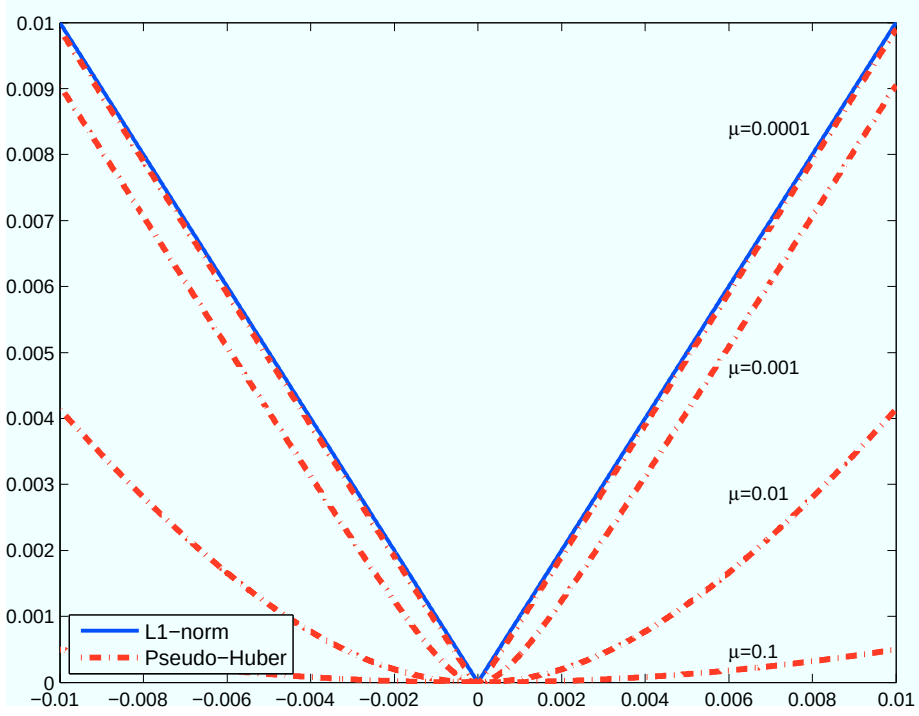
Second-order derivative???

Replace ℓ_1-norm with pseudo-Huber function

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

Smoothing





Empirical observation: the error in the reconstructed components of the signal is $\mathcal{O}(\mu)$.

Newton-Type methods

Two classes of second-order methods

- Inexact Newton methods with Hessian modification: i.e. modified Newton-CG.
  Directions: solve approximately systems

$$(\nabla^2 f_c^\mu(x) + \rho I)d = -\nabla f_c^\mu(x).$$

  Efficient: when the eigenvalues of $\nabla^2 f_c^\mu(x) + \rho I$ have some clustering.

- Quasi Newton methods: i.e. Limited-memory BFGS.
  Directions:

$$d = -B(x)\nabla f_c^\mu(x),$$

  where $B(x)$ is an approximation of the inverse of the Hessian.
  Efficient: when $\nabla^2 f_c^\mu(x) + \rho I$ is well-conditioned and its eigenvalues lack of clustering.

Doubly-Continuation framework

Replace minimize $f_c(x) := \sum_{i=1}^{p} c_i \|W_i x\|_1 + \varphi(x)$

with minimize $f_c^{\mu}(x) := \sum_{i=1}^{p} c_i \psi_{\mu}(W_i x) + \varphi(x)$

Algorithm DC framework

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

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

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

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

using a Newton-Type method with backtracking line-search

Control of spectrum and warm-start

Control of spectrum

- At any x, $0 \prec \nabla^2 f_c^\mu(x) \preceq \left(\frac{1}{\mu}\sum_{i=1}^{p} c_i C(W_i) + \lambda_1\right) I_m$,

  where λ_1 is the largest eigenvalue of $\nabla^2\varphi(x)$

Warm-start of **minimize $\mathbf{f}_c(\mathbf{x})$**

- Zero optimal solution if for an index i

$$c_i \geq \|W_i(W_i^\mathsf{T} W_i)^{-1}\nabla\varphi(0)\|_\infty \text{ for } I_i > m$$

  while the rest c_j, $j \neq i$, regularization parameters are zero

Warm-start of **minimize $\mathbf{f}_c^{\boldsymbol{\mu}}(\mathbf{x})$** (empirical observation)

- Few Newton-type iterations such that $\|\nabla f_c^\mu(x)\| \leq \epsilon$, when

$$c_i \geq \|W_i(W_i^\mathsf{T} W_i)^{-1}\nabla\varphi(0)\|_\infty$$

  while the rest c_j, $j \neq i$, regularization parameters are zero

Iteration complexity of DC framework

- Real, strongly-convex $\varphi(x)$

  K. Fountoulakis and J. Gondzio, *A Second-Order Method for Strongly-Convex L1-Regularization Problems*, Technical Report ERGO-13-011, 2013.

- General convex ℓ_1-regularized problems (discussed in this talk)

  I. Dassios, K. Fountoulakis and J. Gondzio, *Second-order methods for Sparse-Signal Reconstruction*, In preparation.

Signal Processing Problems

We reproduce experiments as given by papers/demos in existing state-of-the-art solver packages.

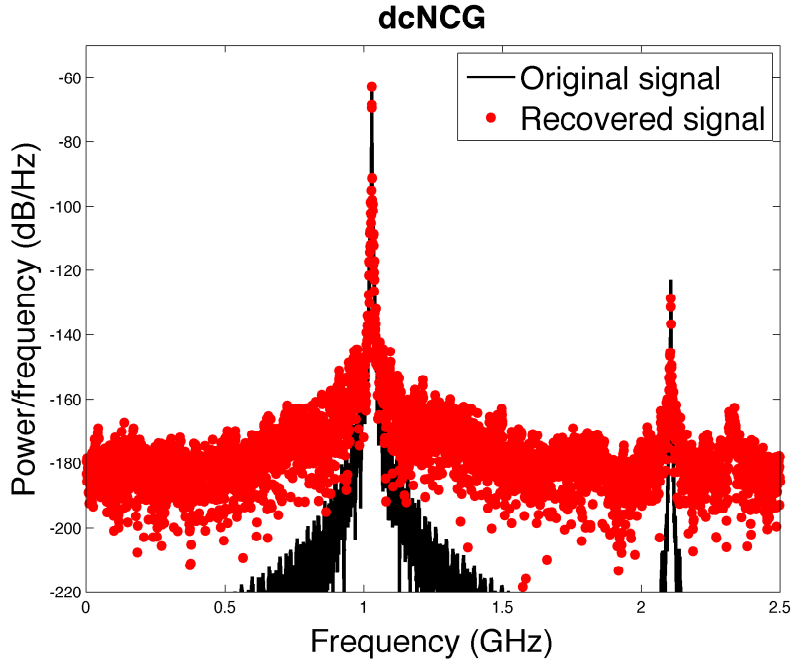
Solvers

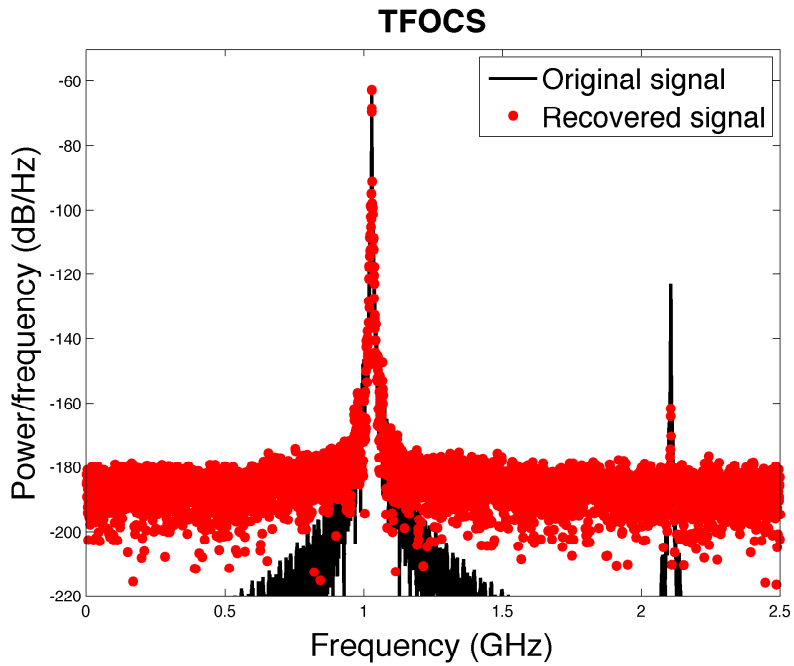
- TFOCS (Templates for First-Order Conic Solvers): Auslender and Teboulle's single-projection method., by S. R. Becker, E. J. Candés and M. C. Grant

ℓ_1-Analysis (recovery of radar pulses)

Info

- W is a Gabor frame
- A is a block diagonal matrix, with ± 1 for entries
- Sub-sampling is $\frac{1}{12}$
- Noise is added so that the small pulse has SNR 0.1 dB
- dcNCG ($\mu = 1.0e\text{-}5$): time=0.3 min., rel. err.=1.56e-3
- TFOCS: time=1.0 min., rel. err.=1.82e-3





Isotropic Total-Variation

Info

- A is a partial Fourier matrix
- Sub-sampling is $\frac{1}{4}$
- SNR 10 dB
- dcNCG ($\mu = 1.0e\text{-}4$):
  time=19.4 sec.,
  PSNR.=17.9 dB
- TFOCS: time=63.2 sec.,
  PSNR=17.8 dB

$$PSNR = 20 \log_{10} \left(\frac{\sqrt{n_1 n_2}}{\|x - \bar{x}\|_F} \right)$$

dcNCG, PSNR is 17.9 dB, CPU time is 19.4

TFOCS, PSNR is 17.8 dB, CPU time is 63.2

Isotropic Total-Variation and ℓ_1-Analysis (denoising)

Info

- Reg.: $\alpha\|Wx\|_1 + \beta\|Dx\|_1$
- W is 9/7 bi-orthogonal wavelet transform
- A is the identity
- SNR 10 dB
- dcNCG ($\mu = 1.0e\text{-}4$): time=6.6 sec., PSNR=27.6 dB
- TFOCS: time=12.7 sec., PSNR=27.5 dB

dcNCG, PSNR 27.6 dB, CPU time is 6.6

TFOCS, PSNR 27.5 dB, CPU time is 12.7

Big-Data and Machine Learning Problems

Sparse Least-Squares

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

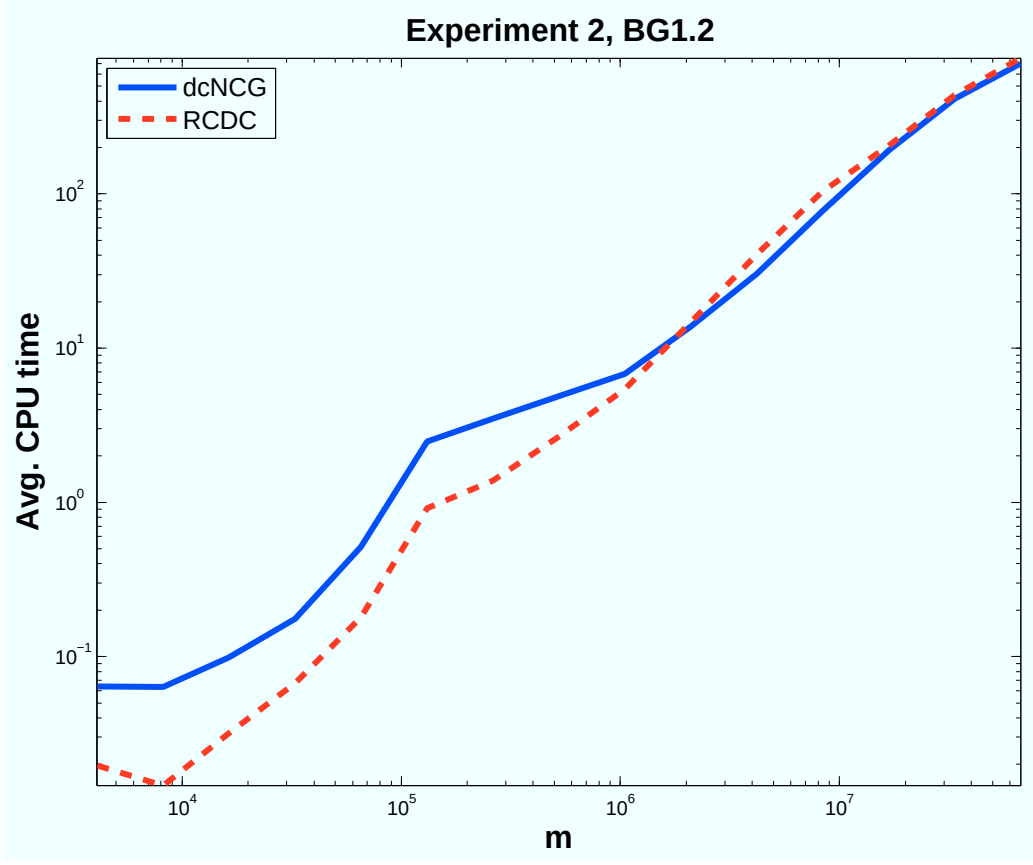
where $x \in \mathbb{R}^m$, $b \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times m}$ with $n \geq m$.

- 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 through multicore 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

- $\frac{\text{mass of } diag(A^\intercal A)}{\text{mass of } A^\intercal A} \approx 99.9\%$
- $n = 2m$
- 20 non-zero elements per column of A
- PCG with $P = diag(A^\intercal A)$
- $\mu = 1.0\text{e}{-12}$



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

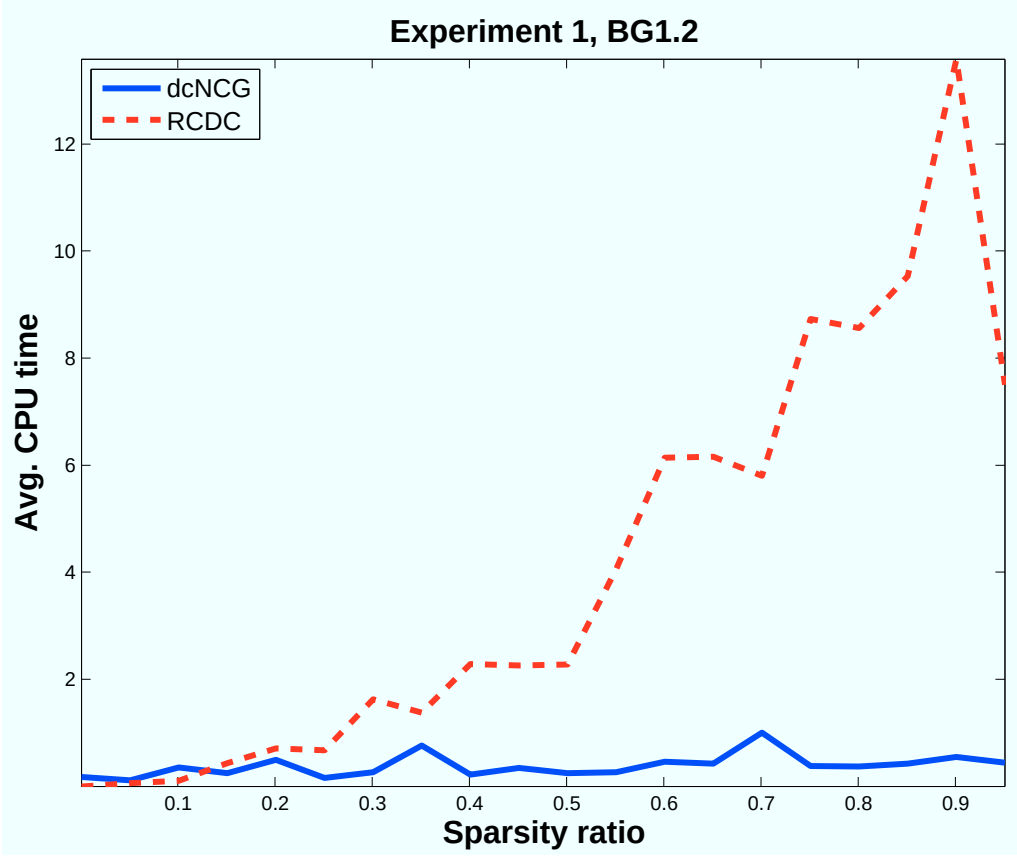
Why is RCDC so fast? Coordinate directions biased with second-order information

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

Sparse Least-Squares: Increasing density

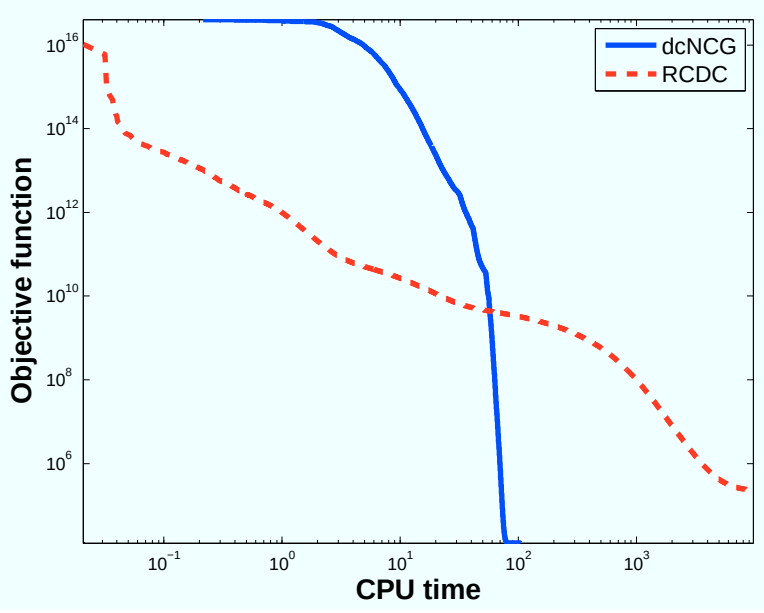
Info

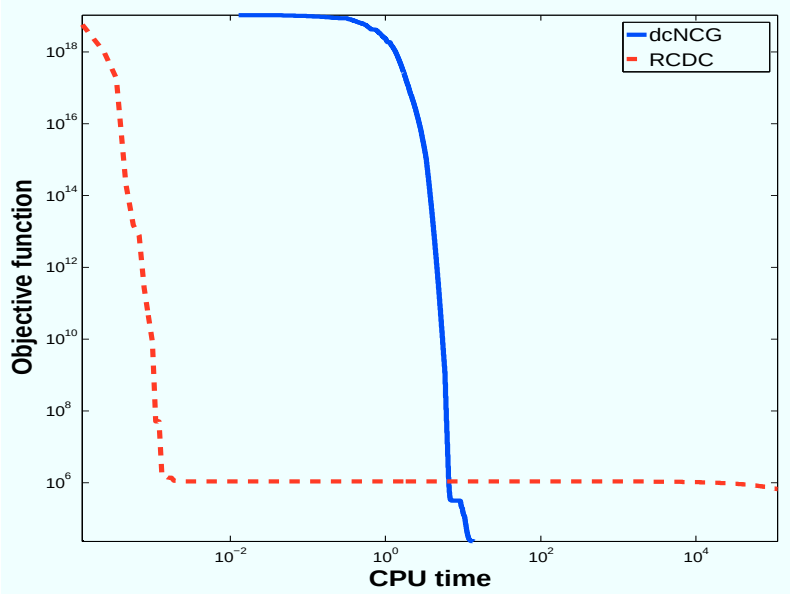
- $\frac{\text{mass of } diag(A^\intercal A)}{\text{mass of } A^\intercal A} \geq 40\%$
- $m = 10^3$, $n = 2m$
- $cond(P^{-1}A^\intercal A) = 10^7$, where $P = diag(A^\intercal A)$
- $\mu = 1.0\text{e}{-12}$
- dcNCG is not affected by the sparsity ratio or any properties of $diag(A^T A)$
- $f(x) - f(x^*) = \mathcal{O}(10^{-6})$



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^{6} and the rest are 10^{-1}.

RCDC was terminated after 10 million iterations $\approx 10k$ seconds and 1 billion iterations $\approx$ 31 hours.

Logistic Regression

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

where $w_i \in \mathbb{R}^n$ 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

Conclusion: Second-order information helps!

Thank you!

Derivatives of pseudo-Huber

For $W : \mathbb{R}^n \to \mathbb{R}^l$

Gradient at x: $\frac{1}{\mu} W^\intercal G W x$, where $G_{ii} = \frac{1}{\sqrt{1 + \frac{[Wx]_i^2}{\mu^2}}} \quad \forall i$

Hessian at x: $\frac{1}{\mu} W^\intercal \tilde{G} W$, where $\tilde{G}_{ii} = G_{ii}^3 \quad \forall i$

For $W : \mathbb{R}^n \to \mathbb{C}^l$

Gradient at x: $\frac{1}{\mu} \mathcal{R}(W^* G W x)$, where $G_{ii} = \frac{1}{\sqrt{1 + \frac{[Wx]_i [\bar{W}x]_i}{\mu^2}}} \quad \forall i$

Hessian at x: $\frac{0.5}{\mu} (\mathcal{R}(W^* Y W) - \mathcal{R}(W^* \tilde{Y} \bar{W}))$,

where $Y_{ii} = G_{ii} + G_{ii}^3$ and $\tilde{Y}_{ii} = \frac{1}{\mu^2} [Wx]_i^2 G_{ii}^3 \quad \forall i$

Termination of N-CG

Newton decrement

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

for $h = d_N$.

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

For d given by CG and $|\nabla f_{\tau^I}^{\mu^I}(x)^T d| = \|d\|_x^2$

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

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

DC itr	NPCG itr	Tot. NPCG itr	a	tau	mu	|d|_x	PCG itr	PCG rs	nMatA	|x|<=1.0e-6	|x|<=1.0e-15
0	1	1	1.00e+00	1.07e+00	1.07e+00	4.74e+00	1	2.11e-16	5	0	0
1	1	2	1.00e+00	5.33e-01	5.33e-01	3.95e-01	3	2.95e-02	13	0	0
2	1	3	1.00e+00	2.66e-01	2.66e-01	1.06e+00	5	5.60e-02	25	0	0
3	1	4	1.00e+00	1.33e-01	1.33e-01	1.54e+00	5	9.82e-02	37	0	0
4	1	5	1.00e+00	6.66e-02	6.66e-02	1.08e+00	6	9.28e-02	51	0	0
5	1	6	1.00e+00	3.33e-02	3.33e-02	5.55e-01	6	9.03e-02	65	0	0
6	1	7	1.00e+00	1.66e-02	1.66e-02	3.05e-01	7	6.54e-02	81	3	0
7	1	8	1.00e+00	8.32e-03	8.32e-03	1.48e-01	7	6.03e-02	97	3	0
8	1	9	1.00e+00	4.16e-03	4.16e-03	7.44e-02	7	6.10e-02	113	0	0
9	1	10	1.00e+00	2.08e-03	2.08e-03	3.89e-02	12	6.94e-03	139	11	0
10	1	11	1.00e+00	1.04e-03	1.04e-03	1.90e-02	12	6.78e-03	165	11	0
11	1	12	1.00e+00	5.20e-04	5.20e-04	9.47e-03	12	7.43e-03	191	28	0
12	1	13	1.00e+00	2.60e-04	2.60e-04	4.73e-03	12	7.43e-03	217	62	0
13	1	14	1.00e+00	1.30e-04	1.30e-04	2.37e-03	12	7.51e-03	243	123	0
14	1	15	1.00e+00	6.50e-05	6.50e-05	1.19e-03	12	7.38e-03	269	236	0
15	1	16	1.00e+00	3.25e-05	3.25e-05	5.90e-04	12	7.15e-03	295	450	0
16	1	17	1.00e+00	1.63e-05	1.63e-05	2.97e-04	12	7.36e-03	321	929	0
17	1	18	1.00e+00	1.00e-05	8.13e-06	1.31e-04	13	6.30e-03	349	1815	0
18	1	19	1.00e+00	1.00e-05	4.06e-06	6.54e-05	11	6.79e-03	373	3245	0
19	1	20	1.00e+00	1.00e-05	2.03e-06	4.51e-05	12	8.69e-03	399	3979	0
20	1	21	1.00e+00	1.00e-05	1.02e-06	3.28e-05	15	7.24e-03	431	4043	0
21	1	22	1.00e+00	1.00e-05	5.08e-07	2.27e-05	13	7.49e-03	459	4045	0
22	1	23	1.00e+00	1.00e-05	2.54e-07	1.64e-05	13	6.68e-03	487	4045	0
23	1	24	1.00e+00	1.00e-05	1.27e-07	1.13e-05	13	6.28e-03	515	4045	0
24	1	25	1.00e+00	1.00e-05	6.35e-08	8.19e-06	1	5.47e-03	519	4045	0
25	1	26	1.00e+00	1.00e-05	3.17e-08	5.82e-06	1	8.89e-03	523	4045	0
26	1	27	1.00e+00	1.00e-05	1.59e-08	4.08e-06	1	6.03e-03	527	4045	0
27	1	28	1.00e+00	1.00e-05	7.94e-09	2.90e-06	1	3.48e-03	531	4045	0
28	1	29	1.00e+00	1.00e-05	3.97e-09	2.03e-06	1	1.88e-03	535	4045	0
29	1	30	1.00e+00	1.00e-05	1.98e-09	1.42e-06	1	1.01e-03	539	4045	0
30	1	31	1.00e+00	1.00e-05	9.92e-10	1.03e-06	1	5.11e-04	543	4045	0
31	1	32	1.00e+00	1.00e-05	4.96e-10	7.25e-07	1	2.61e-04	547	4045	0
32	1	33	1.00e+00	1.00e-05	2.48e-10	5.12e-07	1	1.28e-04	551	4045	0
33	1	34	1.00e+00	1.00e-05	1.24e-10	3.56e-07	1	6.67e-05	555	4045	1
34	1	35	1.00e+00	1.00e-05	6.20e-11	2.56e-07	1	3.30e-05	559	4045	1
35	1	36	1.00e+00	1.00e-05	3.10e-11	1.80e-07	1	1.69e-05	563	4045	2
36	1	37	1.00e+00	1.00e-05	1.55e-11	1.28e-07	1	8.21e-06	567	4045	3
37	1	38	1.00e+00	1.00e-05	7.75e-12	8.94e-08	1	4.21e-06	571	4045	7
38	1	39	1.00e+00	1.00e-05	3.88e-12	6.43e-08	1	2.06e-06	575	4045	8
39	1	40	1.00e+00	1.00e-05	1.94e-12	4.47e-08	1	1.05e-06	579	4045	10
40	1	41	1.00e+00	1.00e-05	9.69e-13	3.19e-08	1	5.19e-07	583	4045	19
41	1	42	1.00e+00	1.00e-05	4.84e-13	2.23e-08	1	2.65e-07	587	4045	31
42	1	43	1.00e+00	1.00e-05	2.42e-13	1.61e-08	1	1.29e-07	591	4045	61
43	1	44	1.00e+00	1.00e-05	1.21e-13	1.12e-08	1	6.53e-08	595	4045	128
44	1	45	1.00e+00	1.00e-05	6.06e-14	7.96e-09	1	3.20e-08	599	4045	255
45	1	46	1.00e+00	1.00e-05	3.03e-14	5.58e-09	1	1.65e-08	603	4045	484
46	1	47	1.00e+00	1.00e-05	1.51e-14	4.01e-09	1	8.07e-09	607	4045	977
47	1	48	1.00e+00	1.00e-05	7.57e-15	2.80e-09	1	4.12e-09	611	4045	1930
48	1	49	1.00e+00	1.00e-05	3.78e-15	2.00e-09	1	2.01e-09	615	4045	3393
49	1	50	1.00e+00	1.00e-05	1.89e-15	1.39e-09	1	1.03e-09	619	4045	3985
50	1	51	1.00e+00	1.00e-05	1.00e-15	9.11e-10	1	5.76e-10	623	4045	4045

Isotropic Total-Variation, deconvolution problem

Info

- A is a deconvolution operator
- blur uniform 9*9, SNR 40 dB,
- dcQN ($\mu = 1.0e$-15): time=10.91 min., ISNR.=17.4 dB
- TwIST: time=9.48 min., ISNR=17.3 dB

$$ISNR = 10 \log_{10} \left(\frac{\|vec(x - x_n)\|_2}{\|vec(x - \bar{x})\|_2} \right)$$

DCNCG

TwIST