Slide 1

ISMP, Lausanne, August 24-29, 1997

Newton's Method for Large Bound-Constrained Optimization Problems

Chih-Jen Lin and Jorge J. Moré

Slide 2

Outline

$$\min f(x)$$
$$l \le x \le u$$

- Trust region Newton method
- Preconditioned conjugate gradient method
- Incomplete Cholesky factorization with limited memory
- Numerical results
- Conclusions and future research

Slide 3

Trust Region Method : Bound Constraints

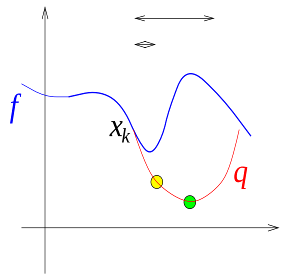
- Compute an approximate solution s_k to the subproblem

$$\min \{q_k(w) : \|w\|_2 \le \Delta_k, \quad l \le x_k + w \le u\},$$

where q_k is the quadratic model.

- Compute the ratio of actual to predicted reduction

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{q_k(0) - q_k(s_k)}$$



- Set $x_{k+1} = \begin{cases} x_k + s_k & \text{if } \rho_k \ge \sigma_0 \\ x_k & \text{if } \rho_k < \sigma_0 \end{cases}$

Slide 4

Main Issues

- Constraint identification
  - Which constraints are active at the solution?
  - The Cauchy point
- Determining the step
  - Global convergence
  - Rates of convergence
  - Degeneracy : $\nabla f(x)_i = 0, x_i = l_i$ or u_i
- Computing the step

Slide 5

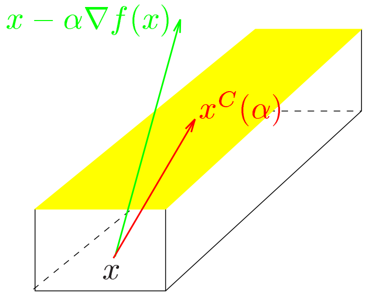
Computing the Cauchy Point

Let $P : \mathbb{R}^n \mapsto \mathbb{R}^n$ be the projection onto the feasible set and define

$$x^C(\alpha) = P[x - \alpha \nabla f(x)].$$

Choose $\alpha > 0$ such that

$$\left\| x^C(\alpha) - x \right\|_2 \le \Delta, \quad q\left(x^C(\alpha)\right) \le q(x) + \mu_0 \nabla q(x)^T \left(x^C(\alpha) - x\right)$$



Slide 6

Solution of the Trust Region Subproblem

Initialize $x = x^C$ and $s = s^C$.

At each minor iteration, compute w such that

$$\min\left\{ q(x+w) : \; w_i = 0, \; i \in \mathcal{A}(x), \; \|w + s\|_2 \le \Delta \right\}$$

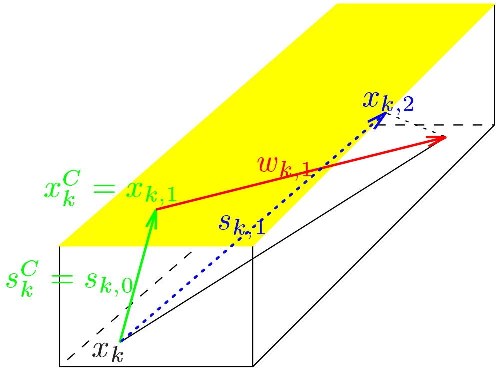
Set $x_+ = P(x + \alpha w)$, where

$$q(x_+) \le q(x) + \mu_0 \nabla q(x)^T (x_+ - x).$$

Caution: The step w may not be a descent direction

Slide 7

Project Gradient and Sufficient Decrease



- The step $w_{k,1}$ from x^C is generally successful
- We can add many constraints at each iteration
- Global and quadratic convergence for degenerate problems.

Slide 8

Solution of the Subproblem

- $\min\{q(x+w) : \; w_i = 0, \; i \in \mathcal{A}(x), \; \|w+s\|_2 \le \Delta\}$ is difficult
- Steihaug [1983] suggested using the PCG method on the model

$$q(w) = g^T w + \frac{1}{2} w^T B w$$

  to generate a step s for the subproblem. B is $\nabla^2 f(x_k)$ or approximation

- The CG algorithm on q generates a sequence $\{w_l\}$ such that

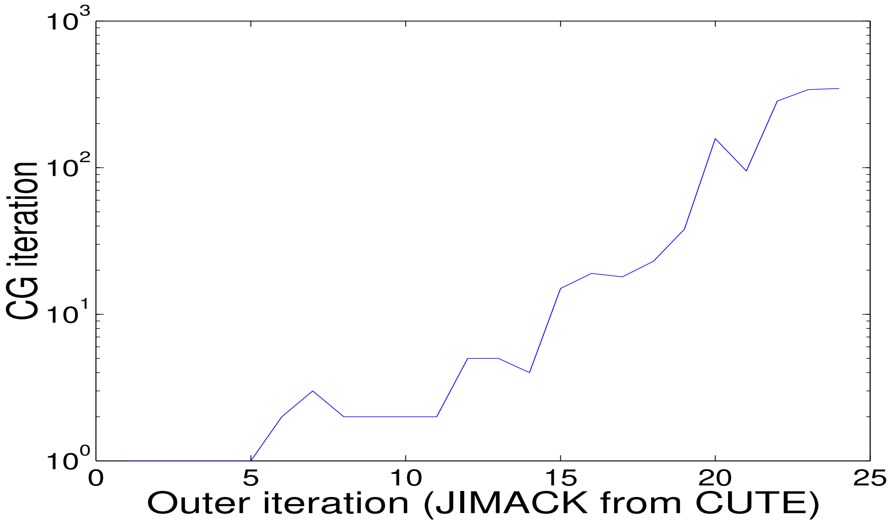
$$w_{l+1} = w_l + \alpha_l p_l$$

  The iterate w_{l+1} is well defined if $p_l^T B p_l > 0$. Moreover,

$$\|w_l\| \le \|w_{l+1}\|$$

Slide 9

Preconditioned Conjugate Gradient Method



- Matrices (P.D. or indefinite) close to singular $\Rightarrow$ more difficult

Slide 10

Incomplete Cholesky Factorization

Saad [1994], Jones and Plassmann [1995]

Let m_j be the number of nonzero off-diagonal elements in the j-th column of the lower triangular part of B.

For $j = 1, 2, \ldots,$

Store the jth column off-diagonal elements to w

Update w by using 1st to $j - 1$ st column of L

Select the largest $m_j + p$ nonzeros in w and store back to L

Slide 11

Computing the ICF

Let B be a symmetric matrix.

Compute $\hat{B} = D^{-1/2} B D^{-1/2}$ where $D = \text{diag}(\|Be_i\|_2)$.

Let μ be a small positive number.

Set $\alpha_0 = 0$ if $\min(b_{i,i}) > 0$; otherwise $\alpha_0 = -\min(b_{i,i}) + \mu$.

For $k = 0, 1, \ldots,$

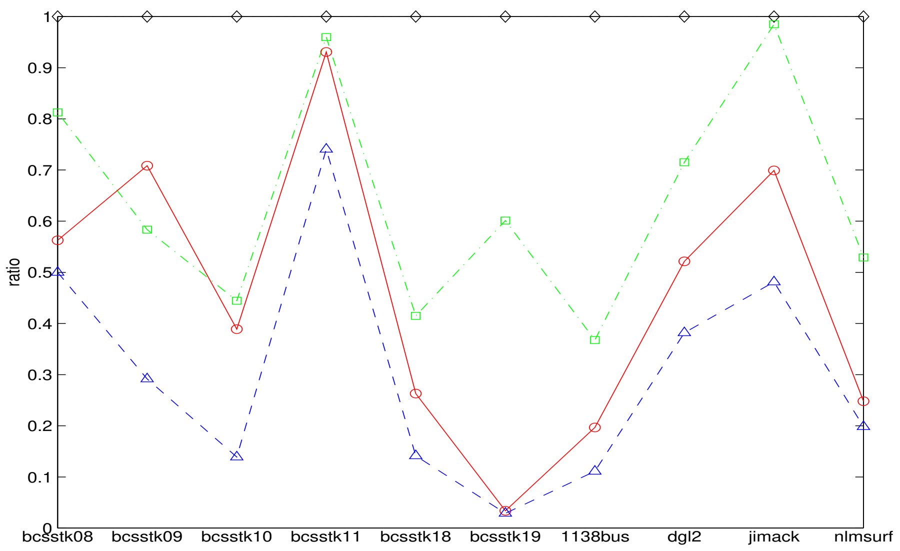
Compute the ICF of $\hat{B}_k = \hat{B} + \alpha_k I$; exit if successful.

Set $\alpha_{k+1} = \max(2\alpha_k, \mu)$

- $\hat{B} + \sqrt{\beta} I$: H-matrix and ICF succeeds, β : max. # of nonzeros in any row of A

Slide 12

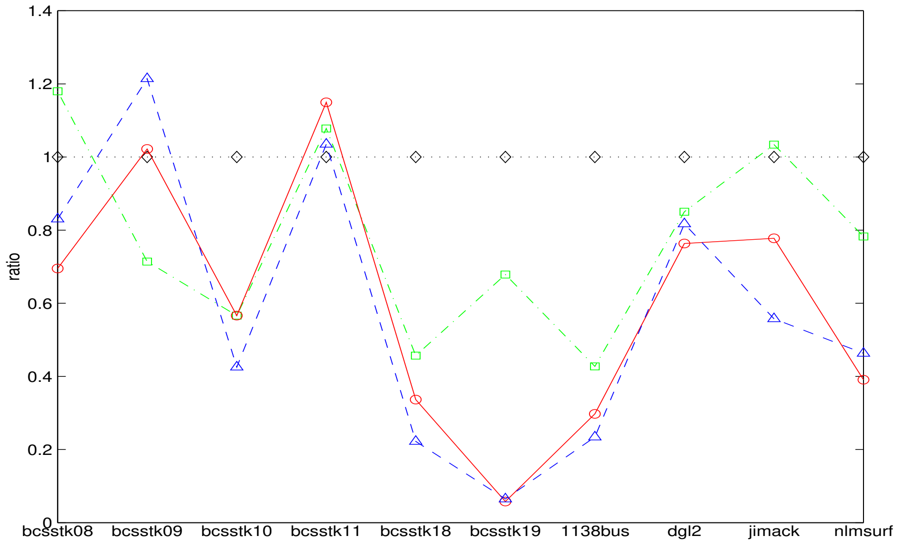
Number of Conjugate Gradient Iterations



($p = 0$ $\Diamond$, $p = 2$ $\square$, $p = 5$ o, $p = 10$, $\triangle$,Harwell-Boeing,Minpack-2,CUTE)

Slide 13

Computational Time and p



$(p = 0\ \Diamond, p = 2\ \square, p = 5\ \text{o}, p = 10, \triangle)$

Slide 14

CG Software Evaluation : CG iteration

	ICF	**ma31 Munksgaard [1980]**			**MATLAB 5.0**	
	$p = 5$	$(10^{-1}, 1)$	$(10^{-3}, 2)$	$(10^{-9}, 4)$	10^{-1}	10^{-3}
bcsstk08	9	36	5	1	52	14
bcsstk09	17	508	84	28	69	6
bcsstk10	14	1086*	11	1	1086*	8
bcsstk11	671	1473*	1473*	1	1473*	1262
bcsstk18	147	11948*	14	219	4893	54
bcsstk19	21	817*	31	2	817*	127
1138bus	23	258	4	1	148	31
dgl2	97	10000*	10000*	10000*	1760	215
jimack	93	1029*	126	2	1029*	128
nlmsurf	30	1024*	17	3	168	12

*: exceed maximal number of iterations

Slide 15

CG Software Evaluation : nnz(L)/nnz(A)

	ICF	ma31			MATLAB 5.0	
	$p = 5$	$(10^{-1}, 2)$	$(10^{-3}, 3)$	$(10^{-9}, 5)$	10^{-1}	10^{-3}
bcsstk08	1.73	1.31	2.31	4.31	0.36	2.88
bcsstk09	1.55	1.22	2.22	4.22	0.44	2.32
bcsstk10	1.46	1.19	2.19	4.19	0.46	1.50
bcsstk11	1.39	1.16	2.16	4.16	0.46	2.36
bcsstk19	2.02	1.43	2.43	4.43	0.43	1.71
bcsstk18	1.56	1.30	2.30	4.30	0.37	1.92
1138bus	2.19	1.88	2.88	4.88	0.95	2.95
dgl2	1.74	1.30	2.30	4.30	0.42	4.52
jimack	1.16	1.07	2.07	4.07	0.37	0.86
nlmsurf	2.00	1.42	2.42	4.42	0.69	3.06

Slide 16

Large Scale Problems : Different Preconditioners

		ICF ($p = 5$)			MATLAB (cholinc)		
Problem	n	nhev	nfev	cg	nhev	nfev	cg
BDEXP	5000	13	14	13	13	14	13
BIGGSB1	1000	504	505	504	504	505	505
BQPGAUSS	2003	10	11	1923	17	18	5369
BRATU1D	1003	5	6	5	5	6	5
CHENHARK	1000	158	159	217	158	159	217
CLPLATEA	5041	14	15	78	14	16	45
JIMACK	3549	21	25	1092	107	117	20702
JNLBRNGA	15625	24	25	81	24	25	81
LINVERSE	1999	16	21	46	68	80	106

Slide 17

Large Scale Problems : Compare with SBMIN

		ICF (p = **5**)			SBMIN		
Problem	n	nhev	nfev	cg	nhev	nfev	cg
BDEXP	5000	13	14	13	10	10	12
BIGGSB1	1000	504	505	504	500	500	499
BQPGAUSS	2003	10	11	1923	8	8	2256
BRATU1D	1003	5	6	5	11	11	7
CHENHARK	1000	158	159	217	225	225	648
CLPLATEA	5041	14	15	78	6	8	319
JIMACK	3549	21	25	1092	52	65	683
JNLBRNGA	15625	24	25	81	21	21	1584
LINVERSE	1999	16	21	46	25	28	1172

- SBMIN : band preconditioner, Conn, Gould and Toint [1992]

Slide 18

Conclusions and Future Research

- Trust region Newton method and project gradient are used. Quadratic convergence is obtained for degenerate problems.
- PCG is used for large scale problems
- A preconditioner with limited memory is proposed and tested
- Extend these results to general nonlinear problems