International Conference
Nonlinear Programming and Variational Inequalities
December 15–18, 1998

Newton's Method for Large Bound-Constrained Optimization Problems

Chih-Jen Lin and Jorge J. Moré
more@mcs.anl.gov
http://www.mcs.anl.gov/~more

ARGONNE NATIONAL LABORATORY • UNIVERSITY OF CHICAGO

NMTRSB

A trust region Newton method for solving large-scale problems

$$\min \{ f(x) : x_l \le x \le x_u \} ,$$

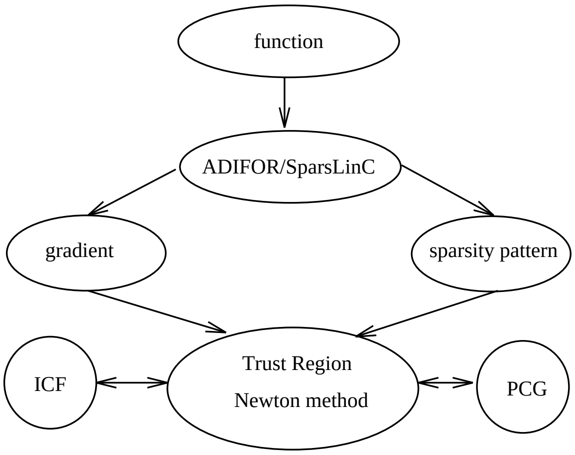
with a preconditioned CG for solving the trust region subproblems.

Advantages

- No assumptions of primal or dual nondegeneracy.
- Global convergence; fast local convergence
- Identification of optimal face in a finite number of iterations.
- Predictable storage requirements
- Reliability

ELSO

Solves large-scale optimization problems while only requiring code for the evaluation of a partially separable function.



Optimization: Background

- The problem

$$\min \{ f(x) : x_l \le x \le x_u \} .$$

- The constraints

$$\Omega = \{ x \in \mathbb{R}^n : x_l \le x \le x_u \} .$$

- The quadratic model

$$q(w) = \nabla f(x)^T w + \tfrac{1}{2} w^T B w.$$

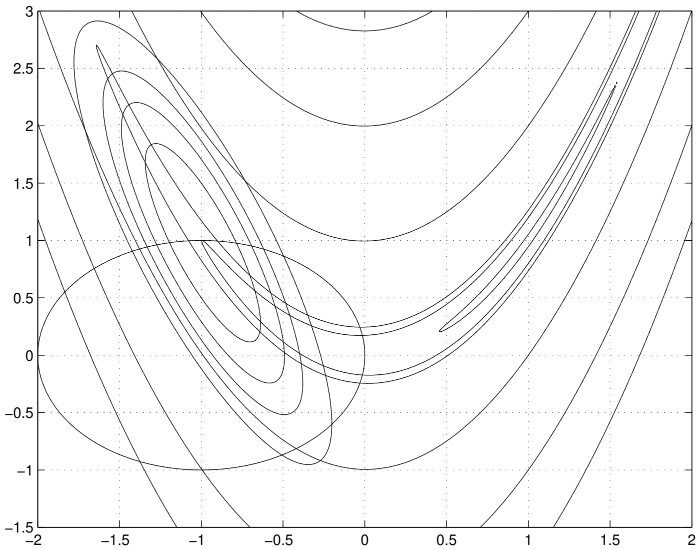
- The actual reduction

$$f(x) - f(x+s).$$

- The predicted reduction

$$q(0) - q(s).$$

Trust Region Methods – A Picture



Contours of $f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1^2)^2$

Trust Region Methods

- Compute a scaling matrix D_k and the *quadratic* model q_k.
- Compute the *Cauchy step* s_k^C.
- Compute a step s_k such that

$$q_k(x_k + s_k) - q(x_k) \le \mu_0 \left\{ q_k(x_k + s_k^C) - q(x_k) \right\},$$

$$\|D_k s_k\| \le \Delta_k, \qquad x_k + s_k \in \Omega.$$

- 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 $x_{k+1} = x_k + s_k$ if $\rho_k \ge \sigma_0$ and $x_{k+1} = x_k$ if $\rho_k < \sigma_0$.

FAQ for Trust Region Algorithms

- What are the convergence results for trust region methods?
- How do we determine the Cauchy step s_k^C?
- How do we determine s_k?
- How do we update Δ_k?
- How do we compute $\nabla f(x_k)$ and $\nabla^2 f(x_k)$?
- How do we compute the sparsity pattern?
- How do we compute the scaling matrix D_k?
  - An incomplete Cholesky factorization.
- How do we determine the step s_k in large-scale problems?
  - Truncated Newton methods.

Projected Gradients

$$\nabla_\Omega f(x) \equiv P_{T(x)}\left[-\nabla f(x)\right] = \operatorname{argmin}\{\|v + \nabla f(x)\| : v \in T(x)\},$$

Properties

- x^* is stationary if and only of $\nabla_\Omega f(x^*) = 0$.
- The mapping $x \mapsto \|\nabla_\Omega f(x)\|$ is lower semicontinuos.
- If $\Omega = [l, u]$, then

$$-[\nabla_\Omega f(x)]_i = \begin{cases} \partial_i f(x) & \text{if } x_i \in (l_i, u_i) \\ \min\{\partial_i f(x), 0\} & \text{if } x_i = l_i \\ \max\{\partial_i f(x), 0\} & \text{if } x_i = u_i \end{cases}$$

Global Convergence of Trust Region Methods

Assumptions

- Ω is a closed, convex set.
- $f : \mathbb{R}^n \mapsto \mathbb{R}$ continuously differentiable on Ω.
- $\{B_k\}$ uniformly bounded.
- x^* is a limit point of $\{x_k\}$

Theorem

There is a subsequence $\{x_{k_i}\}$ such that the steps $\{x_{k_i}\}$ are successful, $\{x_{k_i}^C\}$ converges to x^*, and

$$\lim_{i \to \infty} \|\nabla_\Omega f(x_{k_i}^C)\| = 0.$$

Exposed Faces

We want to show that $x_k \in E\left[-\nabla f(x^*)\right]$ for $k \geq k_0$, where

$$E[d] \equiv \operatorname{argmax}\left\{x \in \Omega : \langle d, x \rangle\right\}.$$

is the face of Ω exposed by $d \in \mathbb{R}^n$.

$-\nabla f(x^*)$

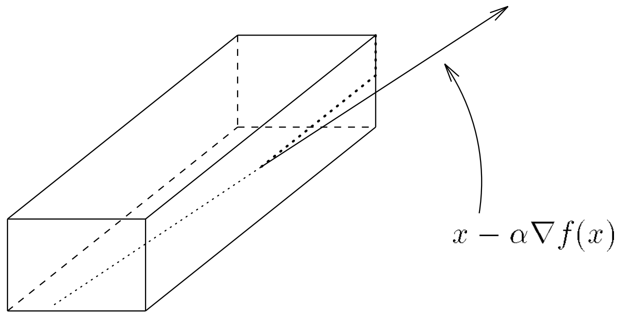
Computing the Cauchy Point

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

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

Choose $\alpha > 0$ such that

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



International Conference on Nonlinear Programming and Variational Inequalities

Exposing Constraints

Assume that the step s_k in the trust region method satisfies

$$\mathcal{A}(x_k^C) \subset \mathcal{A}(x_k + s_k), \qquad k \geq 0.$$

If $\{x_k\}$ converges to x^* then there is an index k_0 such that

$$x_k \in E\left[-\nabla f(x^*)\right], \quad x_k + s_k \in E\left[-\nabla f(x^*)\right], \qquad k \geq k_0.$$

$-\nabla f(x^*)$

Solution of the Trust Region Subproblem

Initialize $x = x^C$.

At each minor iteration, compute w such that

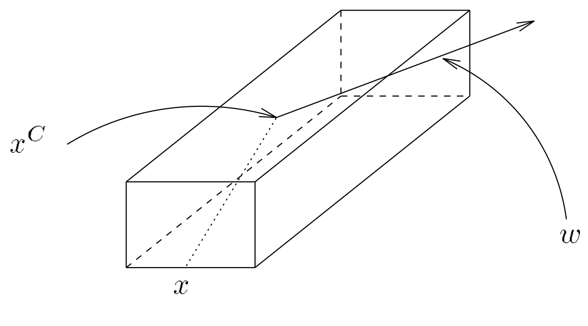
$$\min\Big\{ q(x+w) : \; w_i = 0, \; i \in \mathcal{A}(x), \; \|Dw\|_2 \le \Delta \Big\}$$

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

Projected Searches



- The step from x^C is generally successful
- We can add many constraints at each iteration
- Global and quadratic convergence for degenerate problems.

Local Convergence Results

Regularity means that $\nabla^2 f(x^*)$ is positive definite on the subspace

$$S(x^*) = \{w \in \mathbb{R}^n : w_i = 0,\ i \in \mathcal{B}(x^*)\},$$

of vectors orthogonal to the strictly binding constraints

$$\mathcal{B}(x^*) = \{i \in \mathcal{A}(x^*) : \partial_i f(x^*) \neq 0\}.$$

Theorem

- If $\{x_k\}$ has a limit point x^* and x^* is regular, then $\{x_k\}$ converges to x^*
- If $\{x_k\}$ converges to x^* and x^* is regular, then $\{x_k\}$ converges Q-superlinearly.

The Competition

- Gradient projection, full Newton steps, ϵ-active constraints

  Bertsekas [1982], Gafni and Bertsekas [1984], Schwartz and Polak [1997], Soares, Júdice, and Facchnei [1998].
- Active sets, full Newton steps

  Gay [1984], Forsgren and Murray [1997].
- Gradient projection, truncated Newton steps.

  Conn, Gould, Toint [1988,1989], Lescrenier [1991].
- Interior point, full Newton steps.

  Coleman and Li [1996,1997], Dennis, Heinkenschloss, Vicente [1997], Heinkenschloss, Ulbrich, Ulbrich [1997].

Solving the Trust Region Subproblem

$$\min \{q(w) : \|Dw\| \le \Delta\}, \qquad q(w) = g^T w + \tfrac{1}{2} w^T B w$$

Desirable Properties

- Robust.
- Iterative scheme.
- Predictable storage requirements.
- Termination in a finite number of steps.

Scaling and Incomplete Cholesky Factorizations (ICF)

Original Subproblem

$$\min \{q(w) : \|Dw\| \le \Delta\}, \qquad q(w) = g^T w + \tfrac{1}{2} w^T B w$$

Scaled Subproblem

$$\min \{q_s(v) : \|v\| \le \Delta\}$$

$$q_s(v) = q\left(D^{-1} v\right) = g_s^T v + \tfrac{1}{2} v^T B_s v$$

$$g_s = D^{-T} g, \quad B_s = D^{-T} B D^{-1}$$

The strategy is to choose D so that B_s has clustered eigenvalues

$$B \approx L L^T \Longrightarrow D = L^T$$

Computing the Incomplete Cholesky Factorization

We compute l_j by matching elements in

$$A_j \equiv A - \sum_{k=1}^{j-1} l_k l_k^T = l_j l_j^T + \sum_{k=j+1}^{n} l_k l_k^T .$$

The crucial relations are

$$l_{jj}^2 = a_{jj}^{(j)}, \qquad l_{jj} l_{ij} = a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk}, \quad i > j$$

$$a_{ii}^{(j+1)} = a_{ii}^{(j)} - l_{ij}^2, \qquad i > j.$$

We retain the largest $n_j + p$ elements in l_j, where n_j is the number of nonzero elements in the strict lower triangular part of the original matrix A.

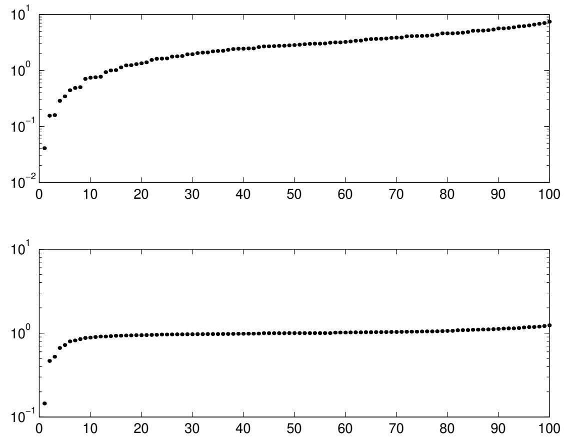
International Conference on Nonlinear Programming and Variational Inequalities

Computing the Incomplete Cholesky Factorization

Let A be a symmetric matrix.

- Choose $\alpha_S > 0$ and $p \geq 0$.
- Compute $\widehat{A} = D^{-1/2} A D^{-1/2}$ where $D = \mathrm{diag}(\|Ae_i\|_2)$.
- Set $\alpha_0 = 0$ if $\min(\widehat{a}_{ii}) > 0$; otherwise $\alpha_0 = -\min(\widehat{a}_{ii}) + \alpha_S$.
- For $k = 0, 1, \ldots,$
  - Compute the ICF of $\widehat{A}_k = \widehat{A} + \alpha_k I$; exit if successful.
  - Set $\alpha_{k+1} = \max(2\alpha_k, \alpha_S)$

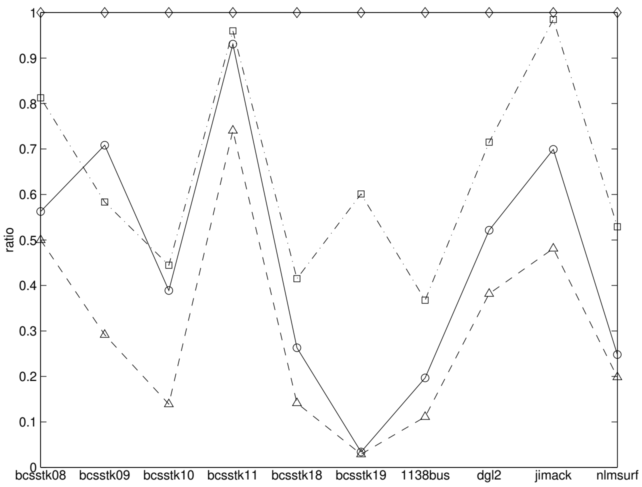
Clustering of Eigenvalues



Eigenvalues of A and $L^{-1}AL^{-T}$ for the msa problem

International Conference on Nonlinear Programming and Variational Inequalities

Numerical Experiments – Dependence on p



Number of conjugate gradient iterates
$\diamond(p=0)$, $\square(p=2)$, $\mathrm{o}(p=5)$, $\triangle(p=10)$

Large Scale Optimization Problems

- Elastic-Plastic Torsion
- Pressure Distribution in a Journal Bearing
- Minimal Surfaces
- Steady-State Combustion
- Homogeneous Superconductors: 2-D Ginzburg–Landau

Properties

- Large-scale problems ($n \geq 10,000$)
- Model problems from application areas.
- Graphical solutions are provided.
- Difficulty depends on problem parameters.
- The function $f : \mathbb{R}^n \mapsto \mathbb{R}$ and the gradient ∇f can be computed in order n operations. The Hessian matrix is sparse.

The Ginzburg-Landau Model for Superconductivity

$$\min\{f(v,A) : v \in K_1,\ A \in K_2\},$$

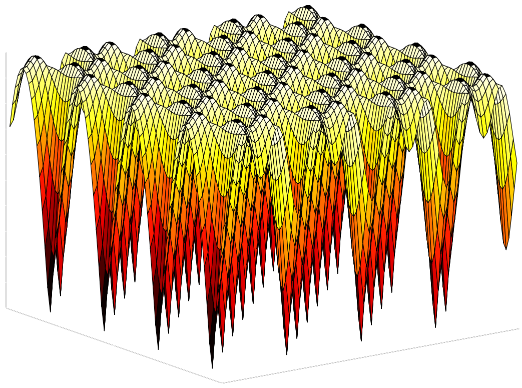
$$f(v,A) = \int_{\mathcal{D}} \left\{ -|v(x)|^2 + \tfrac{1}{2}|v(x)|^4 + \phi(v,A) \right\} \, dx$$

$$\phi(v,A) = \|[\nabla - iA(x)]\, v(x)\|^2 + \kappa^2 \, \|(\nabla \times A)(x)\|^2$$

$$v : \mathbb{R}^2 \to \mathbb{C} \text{ is the order parameter}$$

$$A : \mathbb{R}^2 \to \mathbb{R}^2 \text{ is the vector potential}$$

The Ginzburg-Landau Model for Superconductivity



Surface plot of v for $\kappa = 5$

Computational Experiments – Performance

What parameters impact on performance?

- Number of variables
- Number of minimizers
- Termination accuracy
- Starting point
- Number of active variables at the solution
- Number of negative eigenvalues in the Hessian
- Ill-conditioning of the Hessian
- Sensitivity (nonlinearity) of the problem

Computational Experiments – Performance

What variables measure performance?

- Number of iterations
- Number of function and gradient evaluations
- Number of Hessian evaluations
- Number of conjugate gradient iterations
- Memory requirements
- Computing time

Computational Experiments – Performance

CUTE problems

		NMTRSB				LANCELOT			
Problem	n	nh	nf	ncg	time	nh	nf	ncg	time
BDEXP	5000	10	11	9	2.49	10	11	12	3.52
CVXBQP1	10000	1	2	0	0.52	1	2	1	4.67
JNLBRNGA	15625	22	23	29	25.51	21	22	1584	231.63
MCCORMCK	50000	9	11	10	26.48	4	5	4	22.12
NCVXBQP1	10000	2	3	2	0.90	4	5	0	9.04
NOBNDTOR	14884	37	38	72	44.08	36	37	1386	242.67
NONSCOMP	10000	9	10	9	3.57	8	9	8	5.44
OBSTCLBM	15625	7	8	16	10.50	5	6	1378	281.61
TORSION1	14884	38	39	64	40.10	37	38	1148	171.92

Computational Experiments – Performance

$$n = 10,000$$

	NMTRSB				L-BFGS-B	
Problem	nh	nf	ncg	time	nfg	time
EPT1	29	30	96	18.17	466	53.25
EPT2	30	31	61	15.60	445	42.52
EPT3	20	21	31	8.36	229	16.65
PJB1	21	22	43	17.10	739	91.54
PJB2	12	13	30	9.75	549	58.31
PJB3	6	7	19	5.09	2841	291.45
MSA1	9	16	66	14.53	776	102.46
MSA2	22	26	70	25.56	613	78.34
MSA3	18	19	46	19.91	487	62.00
SSC1	4	5	23	6.23	377	61.78
SSC2	5	6	27	7.87	314	51.83
SSC3	5	6	27	7.84	308	50.59
SSC4	5	6	27	7.88	295	48.39
GL2	6	7	245	39.99	4840	680.34