Nonlinear Optimization and Differential Equations

Jorge Nocedal

with Frank Curtis

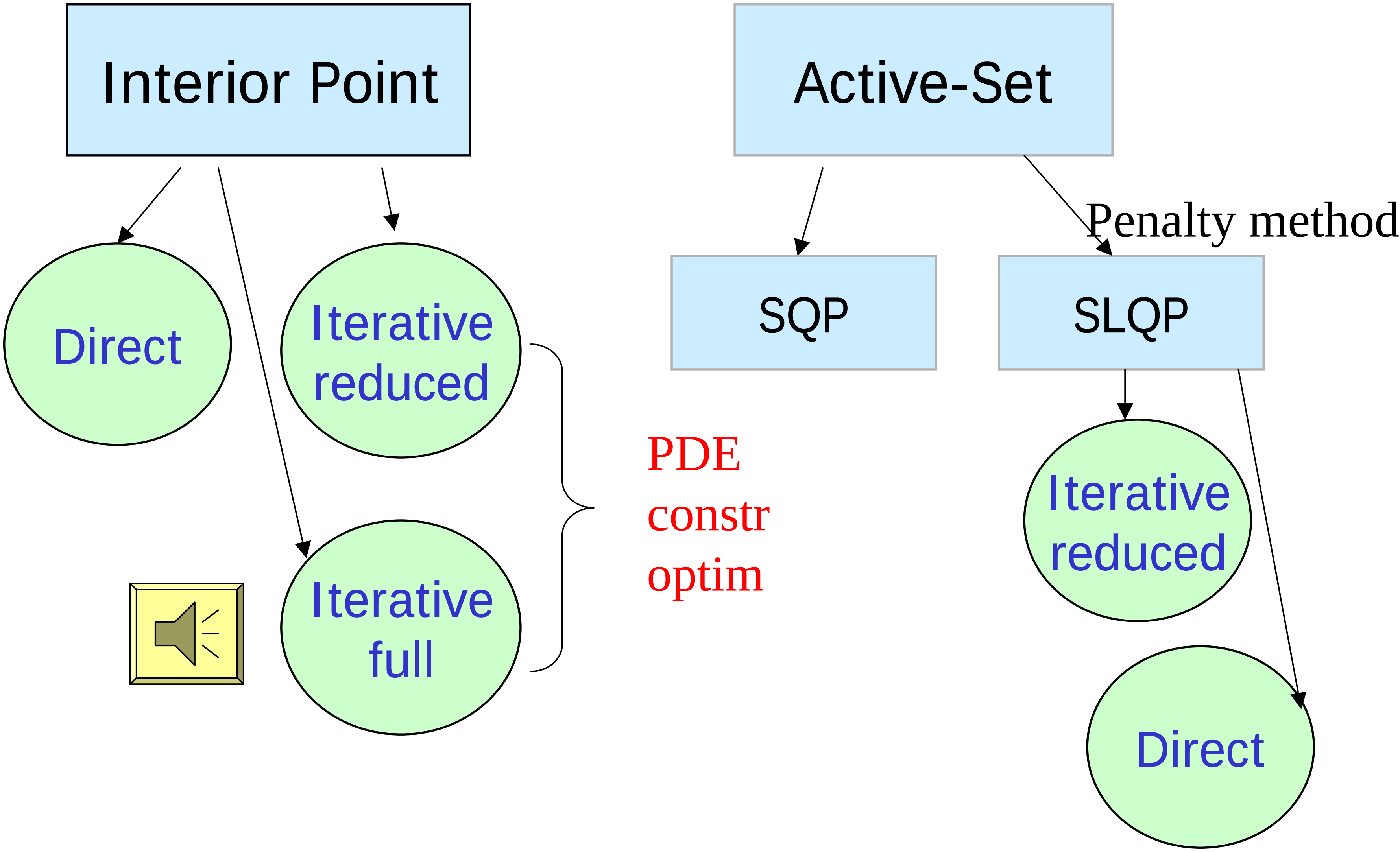
Northwestern University

Livermore, May 2007

Overview

- Review of *some* research contributions by the nonlinear optimization community
- New results: questions in the design of algorithms for PDE-optimization

Nonlinear Optimization, 2007



Looking back at progress made: An Example

- OPF: whose objective is reactive power injection in all network buses as a diagnosis of lack of reactive support in system
- Network represents the Brazilian high voltage generation/transmission system with about 3500 buses and 5000 circuits
- Nonlinear, nonconvex, written in AMPL

Number of variables:	14873
Number of nonlinear equality constraints:	6892
Number of nonzeros in Jacobian:	57971
Number of nonzeros in Hessian:	31501

SQP (SNOPT 7.2):	35 mins
Interior point (KNITRO 5.0):	30 seconds

Why?

- Interior-points much point better in these problems
- Exact second derivatives vs quasi-Newton
- Large reduced space
- Active set method with exact Hessians? See next

Main point: could not get this performance 10 years ago!

In operational models (*n=100,000*) Hessian NOT available
Quasi-Newton needed?

$$\begin{bmatrix} W & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} d \\ \delta \end{bmatrix} = - \begin{bmatrix} \nabla f + A^T \lambda \\ c \end{bmatrix}$$

$$\begin{bmatrix} D & A^T \\ A & 0 \end{bmatrix}$$

Alternative to Quasi-Newton:

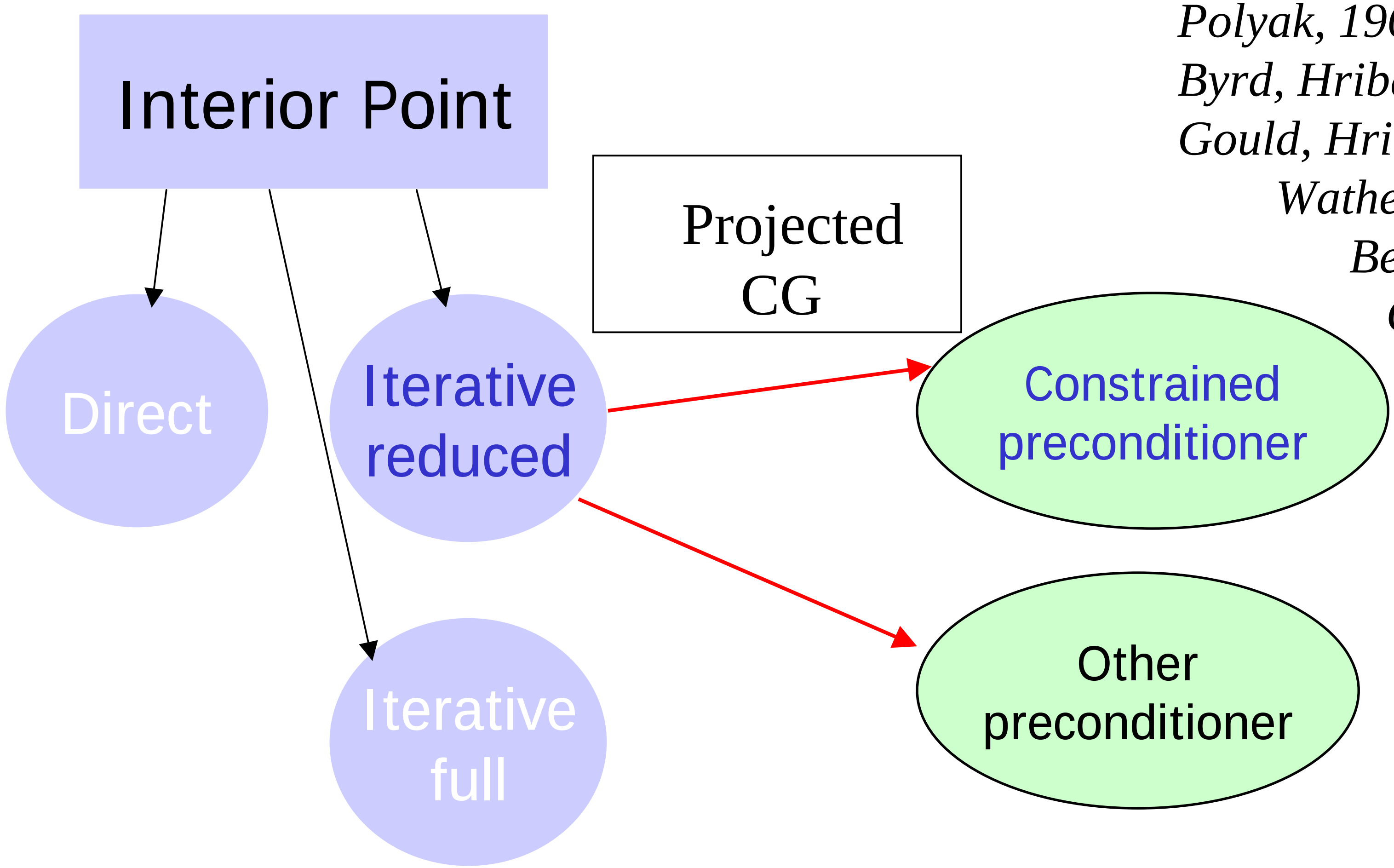
Compute Wd by finite difference of gradients

Iterative Method: reduced space interior point method

Projected CG with constraint preconditioner (to remove barrier ill conditioning)

Software implementation: 2001

Reduced space iterative solve



Polyak, 1969
Byrd, Hribar, N 1999
Gould, Hribar, N. 2001
Wathen, Gould, Benzi, Golub
Gill et al

Performance of various algorithms in KNITRO

Algorithm	Iterations	CPU (secs)
Interior/Default	53	32
Interior/Direct, barrule=4	29	17
Interior/Direct, LBFGS	185	350
Active Set/Hessian	***	> 20 mins (CPLEX)
Interior/CG, FinDiffHess	34	48

Projected CG with constraint preconditioner; NO Hessian

One more comment about recent advances in nonlinear programming

There has been important research done in the last five years on exact penalty functions

$$\phi(x) = f(x) + \pi \| c(x) \|_1$$

- Expands frontiers of NLP, addresses robustness
- Complementarity constraints (theory, algorithms) *Scheel-Scholtes, Anitescu, Ralph, Leyffer, Pang*
- General NLP: new active set methods, *NU, Wright*
- Generall NLP: achieving robustness, *Chen-Goldfarb*
- Dynamic (non-heuristic) rules for updating penalty parameter, *Byrd, Waltz, Nocedal*
- Design of inexact Newton methods for PDE optimization (today!)

Anitescu et al., 00, 04, 07

Benson, Vanderbei, Shanno 04

Byrd, Nocedal, Waltz, 06

Chen and Goldfarb 05, 06

Fletcher and Chin 03

Gould, Orban, Toint 03

Hu and Ralph, 02

Leyffer, Lopez, Nocedal, 04

Scholtes et al 02, 05

Partial list of references

PDE-Optimization

- Inexact Newton Methods for constrained optimization
- Negative Curvature
- Models of Penalty functions

For Simplicity: Equality Constrained Optimization

$$\min_x \quad f(x)$$
$$\text{s.t.} \quad c(x) = 0$$

No. of variables in the millions

Jacobian A not formed but

Av $\quad$ $A^T v$

available

Nonlinear Elimination $\quad$ *Geophysics, Meteorology*

Reduced Space $\quad$ *Biros-Ghattas, Haber,…*

Step-Decomposition $\quad$ *Heinkenschloss, Ridzal*

Full Space (Primal-Dual) SQP

SQP: Approximate Solution of

$$\begin{bmatrix} W & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} d \\ \delta \end{bmatrix} = -\begin{bmatrix} \nabla f + A^T \lambda \\ c \end{bmatrix}$$

$$x_{k+1} = x_k + \alpha_k d \qquad \lambda_{k+1} = \lambda_k + \alpha_k \delta$$

Iterative Methods

- Symmetric QMR
- GMRES
- other …

We impose no structure on the linear solver

When do we stop iterative solver?

Use non-smooth penalty function as a guide!

$$\Phi(x;\pi) = f(x) + \pi \|c(x)\|$$

Negative curvature?

Borrow from trust region methods

retro-1980s

Algorithm: Newton's method

$$\begin{bmatrix} W & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} d \\ \delta \end{bmatrix} = - \begin{bmatrix} \nabla f + A^T \lambda \\ c \end{bmatrix}$$

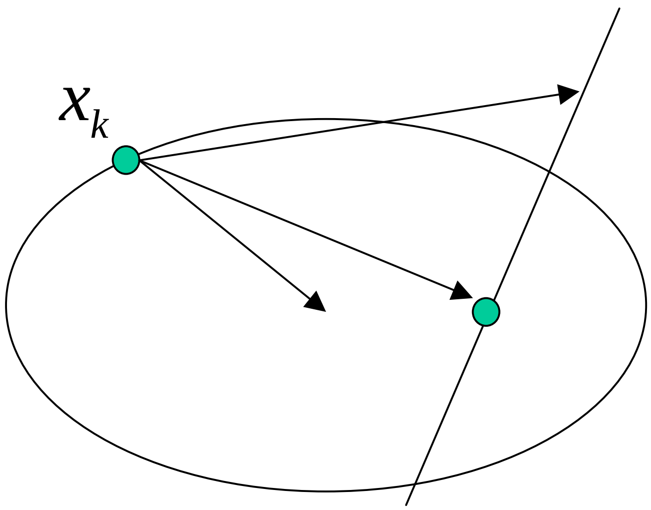
Algorithm: SQP

$$\begin{array}{ll} \min\limits_{d} & \nabla f^T d + \tfrac{1}{2} d^T W d \\ \text{s.t.} & c + Ad = 0 \end{array}$$

Question: can we ensure convergence with a

- step to constraints?
- step to reduce objective?

Preferably both, but if we can't do both?



(Heinkenschloss and Vicente, 2001)

Our approach

$W > 0$

$$\begin{bmatrix} W & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} d \\ \delta \end{bmatrix} = -\begin{bmatrix} \nabla f + A^T \lambda \\ c \end{bmatrix} + \begin{bmatrix} \rho \\ r \end{bmatrix}$$

control residual components separately

Use a model of the merit function

$$m(d) = f + \nabla f^T d + \tfrac{1}{2} d^T W d + \pi(\|c + Ad\|)$$

to determine conditions for ρ and r

Require :

$$\Delta m(d) \geq 0.1\pi \max\{\|c\|, \|r\| - \|c\|\}$$

Model condition implies descent

Instead of imposing descent directly (KNITRO experience)

Algorithm Outline (W>0)

- Iteratively solve

$$\begin{bmatrix} W & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} d \\ \delta \end{bmatrix} = -\begin{bmatrix} g + A^T\lambda \\ c \end{bmatrix} + \begin{bmatrix} \rho \\ r \end{bmatrix}$$

- until

$$\boxed{\begin{array}{l} \|r\| \le \varepsilon \|c\|, \\ \|\rho\| \le \beta \|c\|, \end{array}} \quad \begin{array}{l} 0 < \varepsilon < 1 \\ 0 < \beta \end{array}$$

or

$$\boxed{\begin{array}{l} \|\rho\| \le \varepsilon \|g + A^T\lambda\|, \quad 0 < \varepsilon < 1 \\ \Delta m(d) \ge 0.1\pi \max\{\| c \|, \| r \| - \| c \|\} \end{array}}$$

- Update penalty parameter
- Perform backtracking line search
- Update iterate

If W is positive definite only on the null space of constraints

Only one change: modify the model

$$m(d) = f + \nabla f^T d + \frac{\omega}{2} d^T W d + \pi(\| c + Ad \|)$$

$$\omega = \begin{cases} 1 & if \quad d^T W d \geq \theta \| d \|^2 \\ 0 & \quad \text{otherwise} \end{cases}$$

- linear model for negative curvature direction

Global Convergence

Byrd, Curtis, N., 2006

$$\lim_{k \to \infty} \|c_k\| = 0$$

$$\lim_{k \to \infty} \|g_k + A_k^T \lambda_k\| = 0$$

W is positive definite only on the null space of constraints

Negative curvature

W has negative eigenvalues on null space of A

$$\begin{bmatrix} W & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} d \\ \delta \end{bmatrix} = - \begin{bmatrix} \nabla f + A^T \lambda \\ c \end{bmatrix} + \begin{bmatrix} \rho \\ r \end{bmatrix}$$

Our approach

1. First identify conditions under which the step *d* is acceptable even if negative curvature is present
2. Introduce a modification of *W* if conditions cannot be fulfilled

$$(W + \gamma I)$$

Algorithm Sequential Model Reduction

$$\begin{bmatrix} W+\gamma I & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} d \\ \delta \end{bmatrix} = -\begin{bmatrix} g + A^T\lambda \\ c \end{bmatrix} + \begin{bmatrix} \rho \\ r \end{bmatrix}$$

Repeat until convergence
  Set $\gamma = 0$
  repeat
    Apply 1 or more steps of linear solver
    If Test 1 or Test 2 hold ***break***
    else increase γ
  end repeat
  update penalty
  perform backtracking line search
End repeat

Model Reduction Condition

$$\Delta m(d) \geq (1-\omega)\theta \| u \|^2 + 0.1\pi \max\{\| c \|, \| r \| - \| c \|\}$$

$\omega = 0,1$ $\theta =$ small constant

$u =$ approximation of tangential component of step

How?

Test II:

Model reduction condition

plus

$$\|\rho\| \leq \varepsilon \left\| g + A^T \lambda \right\|,$$

as before…

Test I:

$$\boxed{\begin{aligned} \|r\| &\le \varepsilon \|c\|, \\ \|\rho\| &\le \beta \|c\|, \\ \|u\|^2 &\le \beta \|v\|^2 \end{aligned}}$$

$0 < \varepsilon < 1$

$0 < \beta$

$if \;\; \omega = 0$

How?

$d = u + v$ tangential, normal components

$$\|v\| \ge \|Ad\| / \|d\|$$

$$\|u\|^2 \le \|d\|^2 - \|Ad\|^2 / \|d\|^2$$

Motivation, Justification

Is this all:

- Incremental?
- Recycled?
- Just plain weird?
- .. Or simply wrong headed!

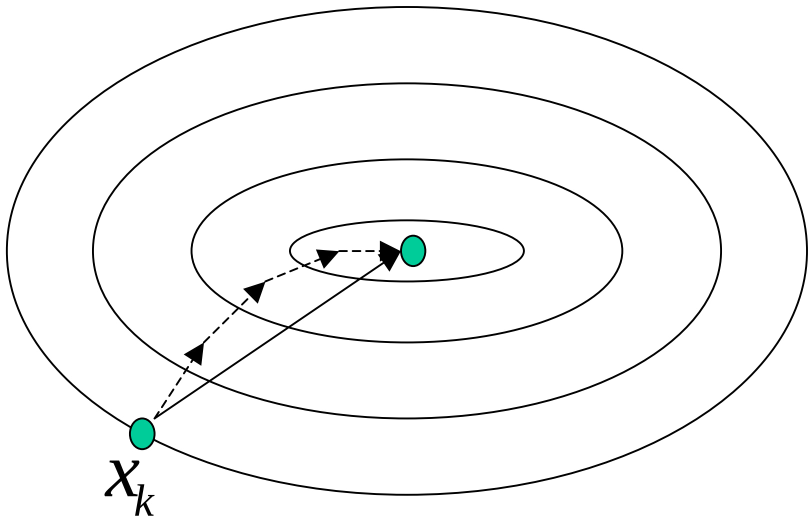
Unconstrained optimization - Inexactness

$$\min_x f(x)$$

Algorithm: Inexact Newton method (CG)

$$\nabla^2 f(x_k) d_k = -\nabla f(x_k)$$

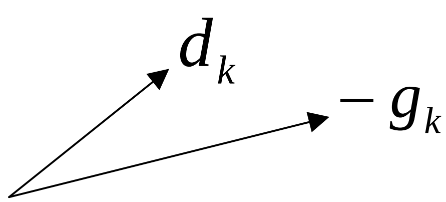
Choosing ***any*** intermediate step ensures global convergence; sufficient (Cauchy) reduction in model



Suppose we cannot use CG…

Negative curvature:

angle condition:



If $\nabla^2 f_k > 0$ (positive def) inexactness condition:

$$\boxed{\nabla^2 f_k d = -\nabla f_k + \rho \qquad \| \rho_k \| \leq \varepsilon \| \nabla f_k \| \qquad \varepsilon < 1}$$

But this condition does not imply descent if *f*. Define model

$$m(d) = f + \nabla f^T d + d^T W d$$

and require

$$\boxed{\Delta m(d) = m(0) - m(d) \geq 0}$$

$$-\nabla f^T d - \frac{1}{2} d^T W d \geq 0 \qquad \text{Easy}$$

Unconstrained optimization: Negative curvature

Exact Newton Method

$$\nabla^2 f(x_k) d_k = -\nabla f(x_k)$$

$$(\nabla^2 f(x_k) + \gamma I) d_k = -\nabla f(x_k)$$

If Hessian not positive def:
- modified Cholesky
- shift modification

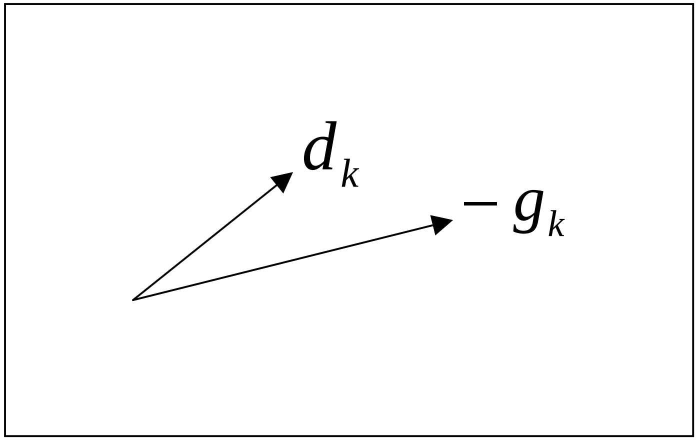
That is, convex models

or

Trust region approach

Why modify so quickly? Step could point downhill (toward saddle point)

Crucial question for inexact Newton case



Unconstrained optimization: Negative Curvature

$$\nabla^2 f_k d = -\nabla f \qquad \text{Newton step}$$

Simple idea: step is acceptable if $\boxed{d^T \nabla^2 f d \geq \theta \| d \|^2}$ (*)

For scale invariance choose $\theta = 10^{-8} \| W \|$

• We prefer this to an angle test, which is not practical in the constrained setting

• Express (*) using a model

$$m(d) = f + \nabla f^T d + \frac{\omega}{2} d^T W d \qquad \omega = 0,1$$

Unconstrained optimization: Model Reduction

Focus on model reduction, not spectrum

Form of model depends on the step d

The new model for Newton's method

$$\boxed{m(d) = f + \nabla f^T d + \frac{\omega}{2} d^T W d}$$

$$\omega = \begin{cases} 1 & \text{if} \quad d^T W d \ge \theta \| d \|^2 \\ 0 & \text{otherwise} \end{cases}$$

Model Reduction Condition: Step d is acceptable if

$$\boxed{\Delta m(d) = m(0) - m(d) \ge (1-\omega)\theta \| d \|^2}$$

If not acceptable, modify W, or add trust region, or…

Model reduction condition ensures descent

A Numerical Test:

Exact Newton step, negative curvature

Algorithm I

$$(\nabla^2 f(x_k) + \gamma I) d_k = -\nabla f(x_k) \qquad \text{(loop)}$$

with $\gamma > 0$ whenever $\text{neig}(\nabla^2 f) > 0$ (margin)

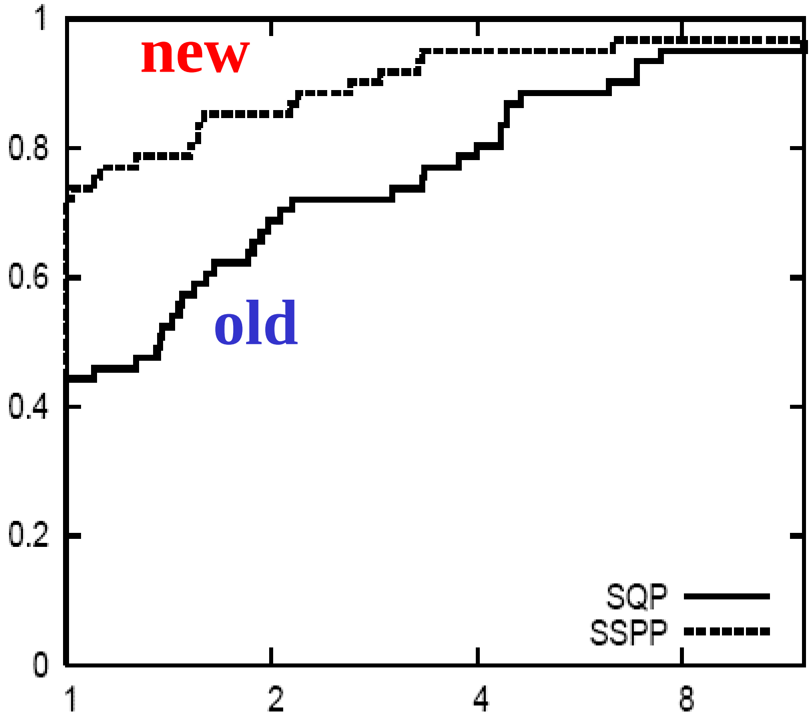
Perform line search

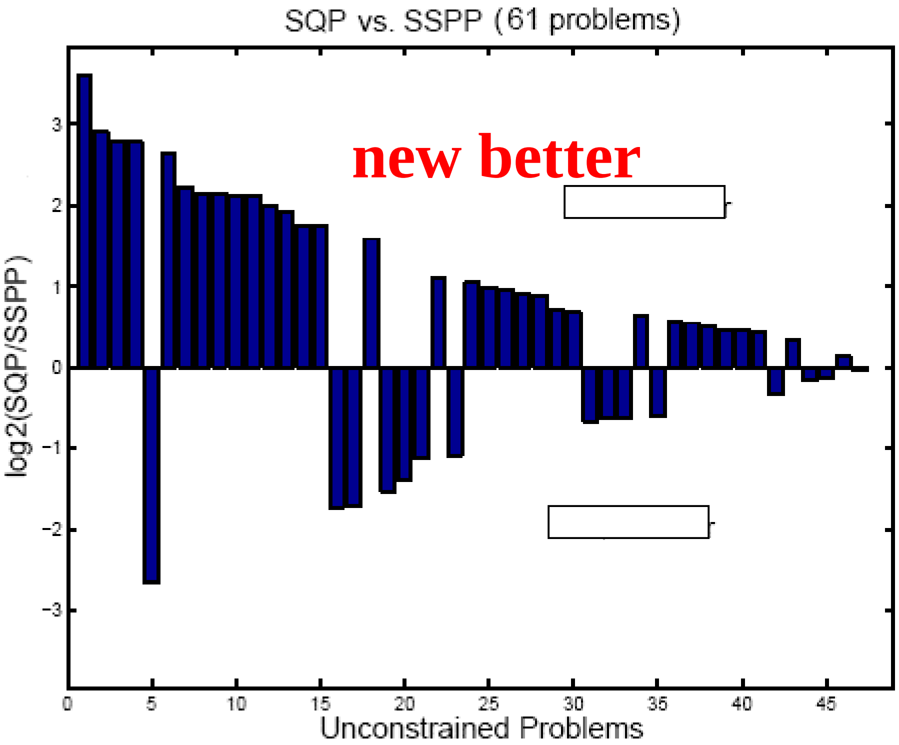
Algorithm II

Shift only if model decrease $\Delta m(d) \geq (1-\omega) 10^{-6} \| d \|^2$

condition does not hold (loop)

Number of factorizations (number of iterations similar for both approaches)
71 problems CUTEr, COPS





Performance profiles for matrix factorizations

Repeat experiments with **iterative solution** of Newton equations
Inertia information not available

$$d^T \nabla^2 f d \geq \theta \| d \|^2$$

Use QMR and SYMQMR

Clear-cut advantage of model reduction approach

Thanks to:

Richard Byrd

Eldad Haber

Richard Waltz

Todd Plantenga

Constrained Optimization

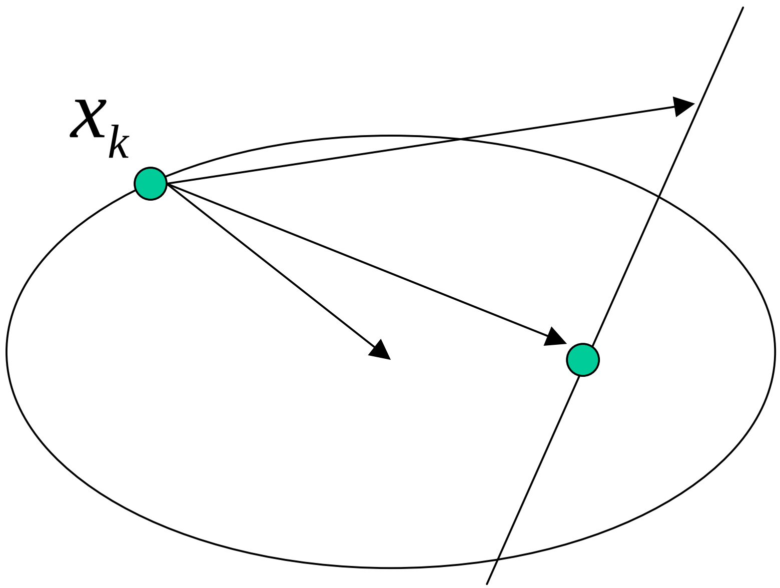
$W > 0 \quad A$ full rank

Algorithm: Newton's method

$$\begin{bmatrix} W & A^T \\ A & 0 \end{bmatrix}\begin{bmatrix} d \\ \delta \end{bmatrix} = -\begin{bmatrix} \nabla f + A^T\lambda \\ c \end{bmatrix}$$

Algorithm: SQP

$$\begin{array}{ll} \min\limits_{d} & \nabla f^T d + \tfrac{1}{2} d^T W d \\ \text{s.t.} & c + Ad = 0 \end{array}$$



Question: can we ensure convergence with a
- step to constraints?
- step to reduce objective?

Preferably both, but if we can't do both?

(Heinkenschloss and Vicente, 2001)

Model Decrease Condition

$$\Phi(x;\pi) = f(x) + \pi \|c(x)\|$$

$$\begin{aligned} \min_d \quad & \nabla f^T d + \tfrac{1}{2} d^T W d \\ \text{s.t.} \quad & c + Ad = 0 \end{aligned}$$

$$m(d) = f + \nabla f^T d + \tfrac{1}{2} d^T W d + \pi(\|c + Ad\|)$$

Quantify reduction obtained from step

$$\begin{aligned} \Delta(d) &= m(0) - m(d) \\ &= -\nabla f^T d - \tfrac{1}{2} d^T W d + \pi(\|c\| - \|c + Ad\|) \end{aligned}$$

cred(d)

Require :

$$\Delta m(d) \geq 0.1\pi \,\|cred\|$$

Algorithm Outline

- Iteratively solve

$$\begin{bmatrix} W & A^T \\ A & 0 \end{bmatrix}\begin{bmatrix} d \\ \delta \end{bmatrix} = -\begin{bmatrix} g + A^T\lambda \\ c \end{bmatrix} + \begin{bmatrix} \rho \\ r \end{bmatrix}$$

- until

$$\|\rho\| \le \varepsilon \|g + A^T\lambda\|, \qquad 0 < \varepsilon < 1$$
$$\Delta m(d) \ge \sigma\pi\|c\|$$

or

$$\|r\| \le \varepsilon\|c\|, \qquad 0 < \varepsilon < 1$$
$$\|\rho\| \le \beta\|c\|, \qquad 0 < \beta$$

–Update penalty parameter
–Perform backtracking line search
–Update iterate