Matrix-free Optimization

Frank E. Curtis
Northwestern University

involving joint work with
Richard H. Byrd, Jorge Nocedal, and Andreas Wächter

ICCOPT, 2007

Outline

Outline

Equality constrained optimization

We consider *very large-scale* problems of the form

$$\min_{x \in \mathbb{R}^n} f(x)$$
$$\text{s.t. } c(x) = 0$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $c : \mathbb{R}^n \to \mathbb{R}^t$ are smooth (nonconvex) functions

- n in the millions (billions?)
- $t \leq n$
- e.g. PDE-constrained optimization

First order optimality

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t. } c(x) = 0$$

We are interested in finding first order optimal points; i.e., defining the Lagrangian function

$$\mathcal{L}(x, \lambda) \triangleq f(x) + \lambda^T c(x)$$

a first order optimal point is one satisfying

$$\nabla \mathcal{L} = \begin{bmatrix} g(x) + A(x)^T \lambda \\ c(x) \end{bmatrix} = 0$$

where $g(x)$ is the gradient of $f(x)$ and $A(x)$ is the Jacobian of $c(x)$

Method of choice: Newton/SQP

$$\min_{x\in\mathbb{R}^n} f(x) \quad \text{s.t. } c(x)=0$$

$$\begin{bmatrix} g(x)+A(x)^T\lambda \\ c(x) \end{bmatrix} = 0$$

A Newton iteration from the point (x_k, λ_k) has the form

$$\begin{bmatrix} W(x_k,\lambda_k) & A(x_k)^T \\ A(x_k) & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \delta_k \end{bmatrix} = -\begin{bmatrix} g(x_k)+A(x_k)^T\lambda_k \\ c(x_k) \end{bmatrix}$$

where $W(x_k,\lambda_k) \approx \nabla^2_{xx}\mathcal{L}(x_k,\lambda_k)$, which in *convex* regions solves the Sequential Quadratic Programming (SQP) subproblem

$$\min_{d\in\mathbb{R}^n} f(x_k)+g(x_k)^T d + \tfrac{1}{2} d^T W(x_k,\lambda_k) d$$
$$\text{s.t. } c(x_k)+A(x_k)d = 0$$

A globalization technique

$$\begin{bmatrix} W_k & A_k^T \\ A_k & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \delta_k \end{bmatrix} = - \begin{bmatrix} g_k + A_k^T \lambda_k \\ c_k \end{bmatrix}$$

$$\min_{d \in \mathbb{R}^n} f(x_k) + g(x_k)^T d + \frac{1}{2} d^T W(x_k, \lambda_k) d$$
$$\text{s.t. } c(x_k) + A(x_k) d = 0$$

Algorithm outline: at (x_k, λ_k)...

- ... evaluate f_k, g_k, c_k, A_k, and W_k
- ... solve the *primal-dual* equations
- ... perform a line search for the merit function

$$\phi(x; \pi) \triangleq f(x) + \pi \|c(x)\|$$

to find $\alpha_k \in (0, 1)$ satisfying the Armijo condition

$$\phi(x_k + \alpha_k d_k; \pi_k) \leq \phi(x_k; \pi_k) + \eta \alpha_k D\phi(d_k; \pi_k)$$

A globalization technique

$$\begin{bmatrix} W_k & A_k^T \\ A_k & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \delta_k \end{bmatrix} = - \begin{bmatrix} g_k + A_k^T \lambda_k \\ c_k \end{bmatrix}$$

$$\begin{aligned} \min_{d \in \mathbb{R}^n} \; & f(x_k) + g(x_k)^T d + \tfrac{1}{2} d^T W(x_k, \lambda_k) d \\ \text{s.t. } & c(x_k) + A(x_k) d = 0 \end{aligned}$$

Algorithm outline: at (x_k, λ_k)...

- ... evaluate f_k, g_k, c_k, A_k, and W_k
- ... solve the *primal-dual* equations
- ... perform a line search for the merit function

$$\phi(x; \pi) \triangleq f(x) + \pi \|c(x)\|$$

to find $\alpha_k \in (0, 1)$ satisfying the Armijo condition

$$\phi(x_k + \alpha_k d_k; \pi_k) \leq \phi(x_k; \pi_k) + \eta \alpha_k D\phi(d_k; \pi_k)$$

Outline

Working with matrices may be impractical

$$\begin{bmatrix} W_k & A_k^T \\ A_k & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \delta_k \end{bmatrix} = - \begin{bmatrix} g_k + A_k^T \lambda_k \\ c_k \end{bmatrix}$$

What if...

- A_k, A_k^T, and W_k cannot be computed explicitly?
- A_k, A_k^T, and W_k cannot be stored?
- the *primal-dual matrix* cannot be factored?
- an iterative method may be more efficient?

If the products $A_k p$, $A_k^T q$, and $W_k y$ can be computed, we have some answers...

Iterative step computations

From now on, let us assume that we have an iterative procedure for solving the primal-dual equations, which during each *inner iteration* yields (d_k, δ_k) solving

$$\begin{bmatrix} W_k & A_k^T \\ A_k & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \delta_k \end{bmatrix} = - \begin{bmatrix} g_k + A_k^T \lambda_k \\ c_k \end{bmatrix} + \begin{bmatrix} \rho_k \\ r_k \end{bmatrix}$$

for the residuals (ρ_k, r_k)

- How can we be sure that a given inexact step is *acceptable*?
- How small do the residuals need to be?

Outline

Problem Statement
- The Optimization Problem
- Computational Challenges

Algorithm Methodology
- Penalty Function Model Reductions
- The Good, The Bad, and The Ugly

Analysis and Experiments
- Overview of Convergence Results
- Numerical Experiments

A naïve approach

$$\begin{bmatrix} W_k & A_k^T \\ A_k & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \delta_k \end{bmatrix} = - \begin{bmatrix} g_k + A_k^T \lambda_k \\ c_k \end{bmatrix} + \begin{bmatrix} \rho_k \\ r_k \end{bmatrix}$$

$$\min_{d \in \mathbb{R}^n} f_k + g_k^T d + \tfrac{1}{2} d^T W_k d \quad \text{s.t. } c_k + A_k d = 0$$

Algorithm outline: at (x_k, λ_k), given $0 < \kappa < 1$...

- ... evaluate f_k, g_k, c_k, $A_k^T \lambda_k$
- ... solve the *primal-dual* equations until $\|(\rho_k, r_k)\| \leq \kappa \|(g_k + A_k^T \lambda_k, c_k)\|$
- ... perform a line search for the merit function

$$\phi(x; \pi) \triangleq f(x) + \pi \|c(x)\|$$

to find $\alpha_k \in (0,1)$ satisfying the Armijo condition

$$\phi(x_k + \alpha_k d_k; \pi_k) \leq \phi(x_k; \pi_k) + \eta \alpha_k D\phi(d_k; \pi_k)$$

A naïve approach

κ	2^{-1}	2^{-5}	2^{-10}
% Solved	45%	80%	86%

Algorithm outline: at (x_k, λ_k), given $0 < \kappa < 1$...

- ... evaluate f_k, g_k, c_k, $A_k^T \lambda_k$
- ... solve the *primal-dual* equations until $\|(\rho_k, r_k)\| \leq \kappa \|(g_k + A_k^T \lambda_k, c_k)\|$
- ... perform a line search for the merit function

$$\phi(x; \pi) \triangleq f(x) + \pi \|c(x)\|$$

to find $\alpha_k \in (0, 1)$ satisfying the Armijo condition

$$\phi(x_k + \alpha_k d_k; \pi_k) \leq \phi(x_k; \pi_k) + \eta \alpha_k \underbrace{D\phi(d_k; \pi_k)}_{>0?}$$

"Fuget-aboud-it"

$$\begin{bmatrix} W_k & A_k^T \\ A_k & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \delta_k \end{bmatrix} = - \begin{bmatrix} g_k + A_k^T \lambda_k \\ c_k \end{bmatrix} + \begin{bmatrix} \rho_k \\ r_k \end{bmatrix}$$

$$\min_{d \in \mathbb{R}^n} \; f_k + g_k^T d + \tfrac{1}{2} d^T W_k d \quad \text{s.t. } c_k + A_k d = 0$$

Take (d_k, δ_k) and...

- ... "forget" about it being an inexact Newton step
- ... "forget" about it being an approximate SQP solution

We want a technique for determining if (d_k, δ_k) is acceptable that...

- ... allows for possibly very inexact solutions to Newton's equations
- ... integrates both step computation and step selection

Central idea: Sufficient Model Reductions

Modern optimization algorithms work with models.

Take the penalty function

$$\phi(x;\pi) \triangleq f(x) + \pi\|c(x)\|$$

and consider the model

$$m_k(d;\pi) \triangleq f_k + g_k^T d + \pi\|c_k + A_k d\|$$

The reduction in m_k attained by d_k is computed easily as

$$\begin{aligned}\Delta m_k(d_k;\pi) &\triangleq m_k(0;\pi) - m_k(d_k;\pi) \\ &= -g_k^T d_k + \pi(\|c_k\| - \|r_k\|)\end{aligned}$$

and yields

$$D\phi(d_k;\pi) \leq -\Delta m_k(d_k;\pi)$$

Outline

Three levels of complexity

- "The Good": the singular values of $\{A_k\}$ are bounded below and away from zero and W_k is positive definite on the null space of A_k for all k [1]
- "The Bad": the singular values of $\{A_k\}$ are bounded below and away from zero and W_k may be indefinite on the null space of A_k for some k [2]
- "The Ugly": A_k may lose rank and W_k may be indefinite [3]

---

[1] R. H. Byrd, F. E. Curtis, and J. Nocedal, "An Inexact SQP Method for Equality Constrained Optimization," submitted to SIAM Journal on Optimization, 2006.

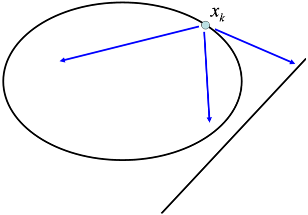
[2] R. H. Byrd, F. E. Curtis, and J. Nocedal, "An Inexact Newton Method for Nonconvex Equality Constrained Optimization," submitted to Mathematical Programming Series A, 2007.

[3] F. E. Curtis, J. Nocedal, and A. Wächter, in preparation.

Main tool: “SMART” Tests

We develop two types of
Sufficient Merit function Approximation Reduction Termination Tests.

Recall the model reduction: $\Delta m_k(d_k;\pi_k) = -g_k^T d_k + \pi_k(\|c_k\| - \|r_k\|)$



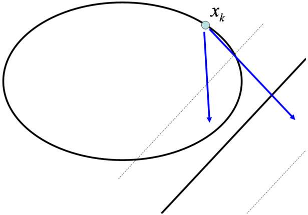
Termination Test I: A sufficient model reduction is attained for π_{k-1} (i.e., the most recent penalty parameter value) and $\sigma \in (0,1)$

$$\Delta m_k(d_k;\pi_{k-1}) \gg 0$$

Main tool: "SMART" Tests

We develop two types of
Sufficient Merit function Approximation Reduction Termination Tests.

Recall the model reduction: $\Delta m_k(d_k;\pi_k) = -g_k^T d_k + \pi_k(\|c_k\| - \|r_k\|)$



Termination Test II: A sufficient reduction in the constraint model is attained for some $\epsilon \in (0,1)$

$$\|r_k\| \leq \epsilon \|c_k\|$$

Step acceptance criteria: "The Good" case

Model Reduction Condition. A step (d_k, δ_k) is acceptable if and only if

$$\Delta m_k(d_k;\pi_k) \geq \max\{\tfrac{1}{2}d_k^T W_k d_k, 0\} + \sigma\pi_k \max\{\|c_k\|, \|c_k + A_k d_k\| - \|c_k\|\}$$

for some $\sigma \in (0,1)$ and an appropriate $\pi_k > 0$.

Termination Test I. For some $\sigma \in (0,1)$ and $\pi_k = \pi_{k-1}$ the Model Reduction Condition is satisfied and for some $\kappa \in (0,1)$ we have

$$\left\| \begin{bmatrix} \rho_k \\ r_k \end{bmatrix} \right\| \leq \kappa \left\| \begin{bmatrix} g_k + A_k^T \lambda_k \\ c_k \end{bmatrix} \right\|$$

Termination Test II. For some $\epsilon \in (0,1)$ and $\beta > 0$ we have

$$\|r_k\| \leq \epsilon\|c_k\| \qquad \text{and} \qquad \|\rho_k\| \leq \beta\|c_k\|$$

and we set

$$\pi_k \geq \frac{g_k^T d_k + \max\{\tfrac{1}{2}d_k^T W_k d_k, 0\}}{(1-\tau)(\|c_k\| - \|r_k\|)} \qquad \text{for } \tau \in (0,1)$$

Inexact SQP with SMART Tests

$$\begin{bmatrix} W_k & A_k^T \\ A_k & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \delta_k \end{bmatrix} = - \begin{bmatrix} g_k + A_k^T \lambda_k \\ c_k \end{bmatrix} + \begin{bmatrix} \rho_k \\ r_k \end{bmatrix}$$

Algorithm outline: at (x_k, λ_k)

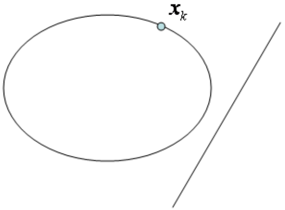
- ... evaluate f_k, g_k, c_k, $A_k^T \lambda_k$
- ... solve the *primal-dual* equations until Termination Test I or II holds
- ... perform a line search for the merit function

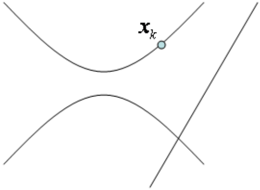
$$\phi(x; \pi) \triangleq f(x) + \pi \|c(x)\|$$

to find $\alpha_k \in (0, 1)$ satisfying the Armijo condition

$$\phi(x_k + \alpha_k d_k; \pi_k) \leq \phi(x_k; \pi_k) + \eta \alpha_k D\phi(d_k; \pi_k)$$

"The Bad" case: handling negative curvature



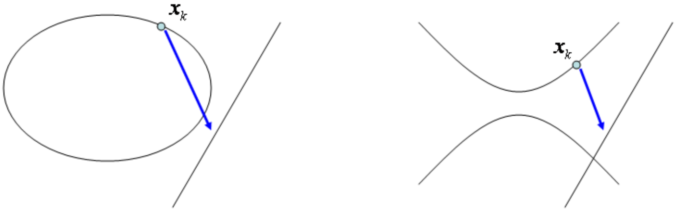


What if the reduced Hessian is not positive definite? We wouldn't know!

$$\begin{bmatrix} W_k & A_k^T \\ A_k & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \delta_k \end{bmatrix} = - \begin{bmatrix} g_k + A_k^T \lambda_k \\ c_k \end{bmatrix} + \begin{bmatrix} \rho_k \\ r_k \end{bmatrix}$$

$$\min_{d \in \mathbb{R}^n} f_k + g_k^T d + \frac{1}{2} d^T W_k d \quad \text{s.t. } c_k + A_k d = 0$$

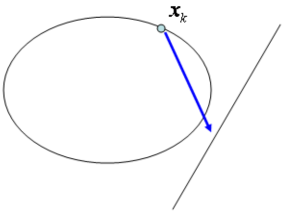
"The Bad" case: handling negative curvature

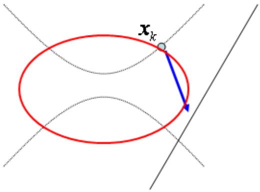


We perform modifications to ensure a reduction in the model, NOT to ensure that the reduced Hessian is positive definite

$$\begin{bmatrix} W_k & A_k^T \\ A_k & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \delta_k \end{bmatrix} = - \begin{bmatrix} g_k + A_k^T \lambda_k \\ c_k \end{bmatrix} + \begin{bmatrix} \rho_k \\ r_k \end{bmatrix}$$

"The Bad" case: handling negative curvature





We want to be sure to recognize a good step even before the model is made convex; i.e., we only modify if

$$\Delta m_k(d_k) \not> 0, \quad \tfrac{1}{2} d_k^T W_k d_k < \theta \|u_k\|^2, \quad \text{and} \quad \|u_k\|^2 > \psi \|v_k\|^2$$

v_k = step to constraints; u_k = step to optimality; $d_k = v_k + u_k$

Step acceptance criteria: "The Bad" case

Model Reduction Condition. A step (d_k, δ_k) is acceptable if and only if

$$\Delta m_k(d_k; \pi_k) \geq \max\{\tfrac{1}{2} d_k^T W_k d_k, \theta\|u_k\|^2\} + \sigma\pi_k \max\{\|c_k\|, \|c_k + A_k d_k\| - \|c_k\|\}$$

for some $\sigma \in (0,1)$, $\theta > 0$, and an appropriate $\pi_k > 0$.

Termination Test I. For some $\sigma \in (0,1)$ and $\pi_k = \pi_{k-1}$ the Model Reduction Condition is satisfied and for some $\kappa \in (0,1)$ we have

$$\left\| \begin{bmatrix} \rho_k \\ r_k \end{bmatrix} \right\| \leq \kappa \left\| \begin{bmatrix} g_k + A_k^T \lambda_k \\ c_k \end{bmatrix} \right\|$$

Termination Test II. For some $\epsilon \in (0,1)$ and $\beta > 0$ we have

$$\|r_k\| \leq \epsilon \|c_k\| \quad \text{and} \quad \|\rho_k\| \leq \beta \|c_k\|$$

and for $\psi > 0$ we have

$$\tfrac{1}{2} d_k^T W_k d_k \geq \theta \|u_k\|^2 \quad \text{or} \quad \psi \|v_k\|^2 \geq \|u_k\|^2$$

and we set

$$\pi_k \geq \frac{g_k^T d_k + \max\{\frac{1}{2} d_k^T W_k d_k, \theta\|d_k\|^2\}}{(1-\tau)(\|c_k\| - \|r_k\|)} \quad \text{for } \tau \in (0,1)$$

Inexact Newton with SMART Tests

$$\begin{bmatrix} W_k & A_k^T \\ A_k & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \delta_k \end{bmatrix} = - \begin{bmatrix} g_k + A_k^T \lambda_k \\ c_k \end{bmatrix} + \begin{bmatrix} \rho_k \\ r_k \end{bmatrix}$$

Algorithm outline: at (x_k, λ_k)

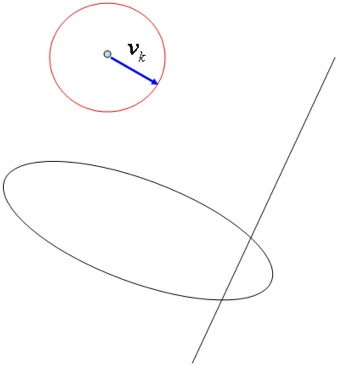
- ... evaluate f_k, g_k, c_k, $A_k^T \lambda_k$
- ... solve the *primal-dual* equations until Termination Test I or II holds, modifying W_k throughout when appropriate
- ... perform a line search for the merit function

$$\phi(x; \pi) \triangleq f(x) + \pi \|c(x)\|$$

to find $\alpha_k \in (0, 1)$ satisfying the Armijo condition

$$\phi(x_k + \alpha_k d_k; \pi_k) \leq \phi(x_k; \pi_k) + \eta \alpha_k D\phi(d_k; \pi_k)$$

"The Ugly" case: handling rank deficiency



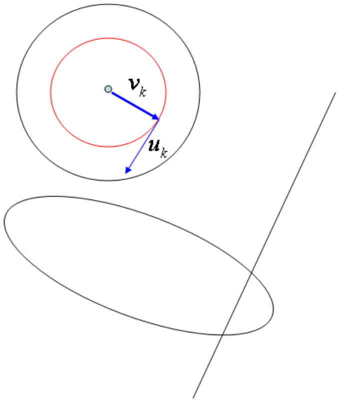
What if the singular values of $\{A_k\}$ are not bounded away from zero?

$$\begin{bmatrix} W_k & A_k^T \\ A_k & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \delta_k \end{bmatrix} = - \begin{bmatrix} g_k + A_k^T \lambda_k \\ c_k \end{bmatrix} + \begin{bmatrix} \rho_k \\ r_k \end{bmatrix}$$

We begin by regularizing the constraint model via a trust region approach

$$\min_{v \in \mathbb{R}^n} \ \tfrac{1}{2}\|c_k + A_k v\|^2$$
$$\text{s.t. } \|v\| \leq \omega \|A_k^T c_k\|$$

"The Ugly" case: handling rank deficiency



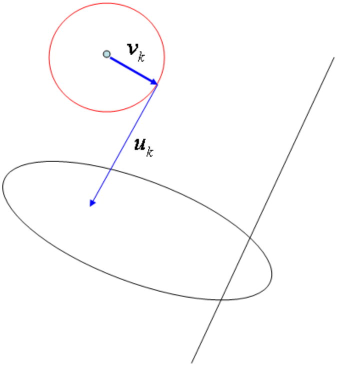
$$\min_{u\in\mathbb{R}^n} \ (g_k + W_k v_k)^T u + \tfrac{1}{2} u^T W_k u$$
$$\text{s.t. } A_k u = 0$$
$$\|u\| \leq \Omega$$

We could consider computing a step toward optimality within a larger trust region, but then we may need

$$Z_k \quad \text{s.t.} \quad A_k Z_k \approx 0$$

or to (approximately) project vectors onto the null space of A_k

"The Ugly" case: handling rank deficiency



$$\min_{u \in \mathbb{R}^n} \; (g_k + W_k v_k)^T u + \tfrac{1}{2} u^T W_k u$$
$$\text{s.t. } A_k u = 0$$

Instead, we turn to the consistent(?) system

$$\begin{bmatrix} W_k & A_k^T \\ A_k & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \delta_k \end{bmatrix} = \begin{bmatrix} -(g_k + A_k^T \lambda_k) \\ A_k v_k \end{bmatrix}$$

or, equivalently,

$$\begin{bmatrix} W_k & A_k^T \\ A_k & 0 \end{bmatrix} \begin{bmatrix} u_k \\ \delta_k \end{bmatrix} = - \begin{bmatrix} g_k + A_k^T \lambda_k + W_k v_k \\ 0 \end{bmatrix}$$

and alter our model reduction criteria

Step acceptance criteria: "The Ugly" case

Model Reduction Condition. A step (d_k, δ_k) is acceptable if and only if

$$\Delta m_k(d_k; \pi_k) \geq \max\{\tfrac{1}{2} d_k^T W_k d_k, \theta \|u_k\|^2\} + \sigma \pi_k (\|c_k\| - \|c_k + A_k v_k\|)$$

for some $\sigma \in (0,1)$, $\theta > 0$, and an appropriate $\pi_k > 0$.

Termination Test I. For some $\sigma \in (0,1)$ and $\pi_k = \pi_{k-1}$ the Model Reduction Condition is satisfied and for some $\kappa \in (0,1)$ we have

$$\left\| \begin{bmatrix} \rho_k \\ r_k \end{bmatrix} \right\| \leq \kappa \left\| \begin{bmatrix} g_k + A_k^T \lambda_k \\ c_k \end{bmatrix} \right\| \quad \text{and} \quad \|A_k u_k\| \leq \zeta$$

Termination Test II. For some $\epsilon \in (0,1)$ and $\beta > 0$ we have

$$\|r_k\| \leq \epsilon \|c_k\| \quad \text{and} \quad \|\rho_k\| \leq \beta \|c_k\|$$

and for $\psi > 0$ we have $\psi \|v_k\|^2 \geq \|u_k\|^2$ or

$$\tfrac{1}{2} d_k^T W_k d_k \geq \theta \|u_k\|^2 \quad \text{and} \quad \|A_k u_k\| \leq \zeta$$

and we set

$$\pi_k \geq \frac{g_k^T d_k + \max\{\frac{1}{2} d_k^T W_k d_k, \theta \|u_k\|^2\}}{(1-\tau)(\|c_k\| - \|r_k\|)} \quad \text{for } \tau \in (0,1)$$

Outline

Brief overview of analysis

- The step length (d_k, v_k, u_k) is explicitly or implicitly controlled...
- The directional derivative of the merit function satisfies

$$D\phi(d_k;\pi_k) \leq -\Delta_k m_k(d_k;\pi_k) \leq -\gamma(\|u_k\|^2 + \|A_k^T c_k\|^2)$$

- In "nice" cases, $\{\pi_k\}$ remains bounded and

$$\lim_{k\to\infty} \left\| \begin{bmatrix} g_k + A_k^T\lambda_k \\ c_k \end{bmatrix} \right\| = 0,$$

but in any case

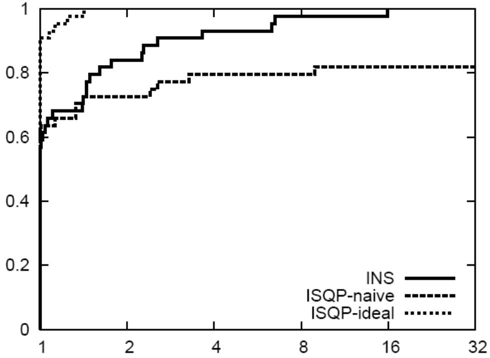
$$\lim_{k\to\infty} \|A_k^T c_k\| = 0$$

Outline

"The Good" case

κ	2^{-1}	2^{-5}	2^{-10}	iSQP
% Solved	45%	80%	86%	100%

"The Bad" case



INS: inexact Newton with SMART Tests

iSQP-naive: inexact SQP with no modifications

iSQP-ideal: inexact SQP with inertia control

Conclusion

We have...

- ... focused on a particular class of problems to which contemporary optimization techniques cannot be applied
- ... considered the fundamental question of how to ensure global convergence via a type of inexact SQP/Newton approach
- ... developed a novel methodology where inexact solutions are appraised based on the reductions obtained in linear models of an exact penalty function
- ... extended the algorithm and analysis for cases involving indefiniteness and rank deficiency