An Inexact Sequential Quadratic Optimization Method for Nonlinear Optimization

Frank E. Curtis, Lehigh University

involving joint work with

Travis Johnson, Northwestern University
Daniel P. Robinson, Johns Hopkins University
Andreas Wächter, Northwestern University

SIAM Conference on Optimization — San Diego, CA

22 May 2014

Outline

Problem formulation

Our goal is to solve a large-scale constrained nonlinear optimization problem:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \ & f(x) \\ \text{s.t. } & c(x) = 0, \ \bar{c}(x) \leq 0. \end{aligned} \tag{NLP}$$

If (NLP) is infeasible, then at least we want to minimize constraint violation:

$$\min_{x \in \mathbb{R}^n} \ v(x), \ \text{where } v(x) := \|c(x)\|_1 + \|[\bar{c}(x)]^+\|_1. \tag{FP}$$

Any feasible point for (NLP) is an optimal solution of (FP).

Algorithm design objectives[1]

Demands on algorithms for large-scale constrained nonlinear optimization:

- Scalable step computation
- Effectively handles negative curvature
- Superlinear convergence in primal-dual space
- Asymptotic monotonicity (consistent progress toward solution)
- Active-set detection and warm-starting

Traditional methods (SQO, IP, AL) only satisfy subsets of these demands.

- "I can solve my problem quickly enough with an IP method using matrix factorizations"... then you're not interested in what I have to say.
- If you can only afford a few NLP iterations, are solving sequences of similar problems, and/or need a method that can converge fast locally, ...

[1] Paraphrased from Zavala and Anitescu (2014)

Sequential quadratic optimization (SQO)

SQO advantages:

- "Parameter free" search direction computation (ideally)
- Strong global convergence properties and behavior
- Active-set identification $\Longrightarrow$ Newton-like local convergence

SQO disadvantages:

- No "best" way to handle negative curvature/inconsistent subproblems
- Quadratic subproblems (QPs) are expensive to solve exactly — not scalable!

Questions

- *Employing scalable QP solvers, can we exploit inexact solves?*
- *From inexactness, do we lose global/local convergence? active set detection?*
- *Is there any benefit in applying a scalable method for the QP subproblems, as opposed to applying a similar scalable method to* (NLP) *directly?*

Sit and solve QP? Or move on nonlinear problem as in `filterSD`; Fletcher (2012)

Inexact SQO vs. inexact SQO

Inexact SQO may mean inexact gradients with (in)exact QP solves:

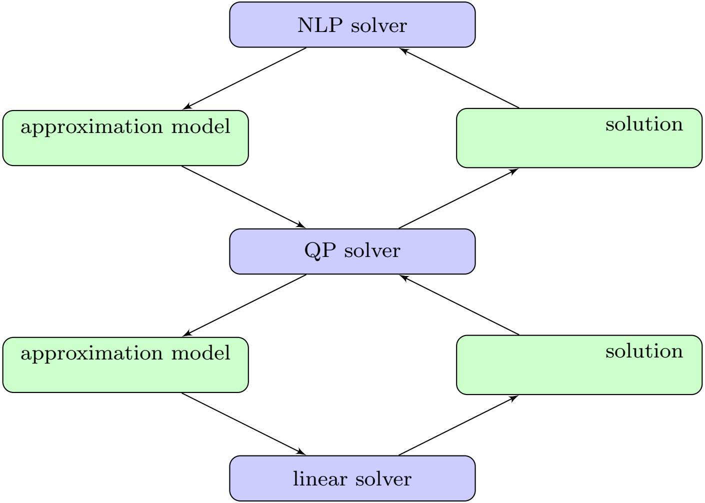
- Dennis, El-Alem, Maciel (1997)
- Diehl, Walther, Bock, Kostina (2010)
- Jäger, Sachs (1997)
- Heinkenschloss, Vicente (2000)
- Walther, Vetukuri, Biegler (2012)

Inexact SQO may also mean exact gradients with inexact QP solves:

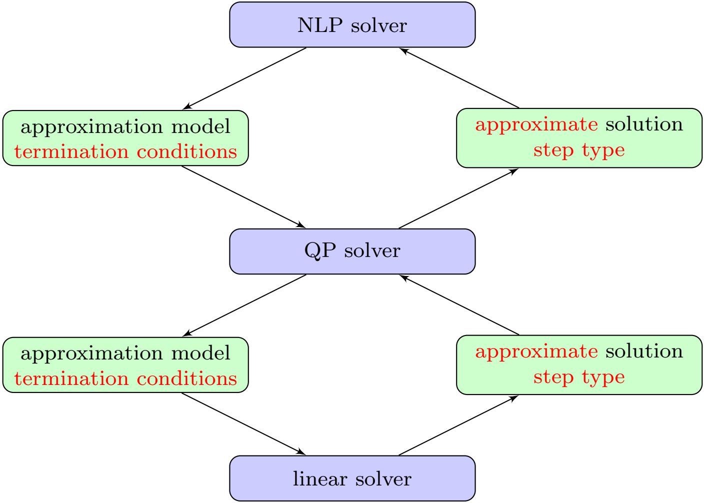
- Byrd, Curtis, Nocedal (2008)
- Izmailov, Solodov (2010)
- Leibfritz, Sachs (1999)
- Morales, Nocedal, Wu (2010)

In this talk, we are interested in the latter type of inexact SQO.

Algorithmic framework: Classic



Algorithmic framework: Inexact



SQO w/ inexact subproblem solves

Contributions:

- Broad, but consistent, framework; Gill, Robinson (2013), pdAL/sSQO
- Implementable termination conditions for inexact QP solves
- No specific QP solver required
- Global convergence guarantees (feasible and infeasible problems)
- Future work: Fast local convergence (feasible and infeasible problems)[2]

Algorithmic features:

- Allows "generic" inexactness in QP solutions
- Convex combination of "optimality" and "feasibility" steps
- Negative curvature handled with dynamic Hessian modifications
- Separate multipliers for (NLP) and (FP)
- Dynamic updates for penalty parameter and Lagrange multipliers

[2] Avoid using "Cauchy points" that only yield minimal progress for global convergence.

Fritz John and penalty functions

(NLP):

$$\min_x \; f(x)$$
$$\text{s.t. } c(x) = 0, \;\; \bar{c}(x) \leq 0$$

(FP):

$$\min_x \; v(x) := \left\| \begin{bmatrix} c(x) \\ [\bar{c}(x)]^+ \end{bmatrix} \right\|_1$$

Define the Fritz John (FJ) function

$$\mathcal{F}(x, y, \bar{y}, \mu) := \mu f(x) + c(x)^T y + \bar{c}(x)^T \bar{y}$$

and the ℓ_1-norm exact penalty function

$$\phi(x, \mu) := \mu f(x) + v(x).$$

$\mu \geq 0$ acts as objective multiplier/penalty parameter.

Optimality conditions

(NLP):

$$\min_x \; f(x)$$
$$\text{s.t. } c(x) = 0, \;\; \bar{c}(x) \le 0$$

(FP):

$$\min_x \; v(x) := \left\| \begin{bmatrix} c(x) \\ [\bar{c}(x)]^+ \end{bmatrix} \right\|_1$$

(PP):

$$\min_x \; \phi(x, \mu) := \mu f(x) + v(x)$$

(FJ):

$$\mathcal{F}(x, y, \bar{y}, \mu) :=$$
$$\mu f(x) + c(x)^T y + \bar{c}(x)^T \bar{y}$$

KKT conditions for (FP) and (PP) expressed with residual

$$\rho(x, y, \bar{y}, \mu) := \begin{bmatrix} \mu g(x) + J(x) y + \bar{J}(x) \bar{y} \\ \min\{[c(x)]^+, e - y\} \\ \min\{[c(x)]^-, e + y\} \\ \min\{[\bar{c}(x)]^+, e - \bar{y}\} \\ \min\{[\bar{c}(x)]^-, \quad \bar{y}\} \end{bmatrix}$$

- FJ point:

$$\rho(x, y, \bar{y}, \mu) = 0, \;\; v(x) = 0, \;\; (y, \bar{y}, \mu) \neq 0$$

- KKT point:

$$\rho(x, y, \bar{y}, \mu) = 0, \;\; v(x) = 0, \;\; \mu > 0$$

- Infeasible stationary point:

$$\rho(x, y, \bar{y}, 0) = 0, \;\; v(x) > 0$$

Penalty function model and QP subproblem

(NLP):

$$\min_x f(x)$$

$$\text{s.t. } c(x) = 0, \;\; \bar{c}(x) \leq 0$$

(FP):

$$\min_x \; v(x) := \left\| \begin{bmatrix} c(x) \\ [\bar{c}(x)]^+ \end{bmatrix} \right\|_1$$

(PP):

$$\min_x \; \phi(x,\mu) := \mu f(x) + v(x)$$

(FJ):

$$\mathcal{F}(x,y,\bar{y},\mu) :=$$

$$\mu f(x) + c(x)^T y + \bar{c}(x)^T \bar{y}$$

KKT residual:

$$\rho(x,y,\bar{y},\mu)$$

Define a local model of $\phi(\cdot,\mu)$ at x_k:

$$l_k(d,\mu) := \mu(f_k + g_k^T d) + \|c_k + J_k^T d\|_1 + \|[\bar{c}_k + \bar{J}_k^T d]^+\|_1$$

Reduction in this model yielded by a given d:

$$\Delta l_k(d,\mu) := \Delta l(0,\mu) - \Delta l(d,\mu)$$

Subproblem of interest:

$$\min_d \; -\Delta l_k(d,\mu) + \tfrac{1}{2} d^T H d \qquad \text{(QP)}$$

$\Delta l_k(d,\mu) > 0$ implies d is a direction of strict descent for $\phi(\cdot,\mu)$ from x_k

Optimality conditions (for QP)

(NLP):

$$\min_x \; f(x)$$
$$\text{s.t. } c(x) = 0, \;\; \bar{c}(x) \leq 0$$

(FP):

$$\min_x \; v(x) := \left\| \begin{bmatrix} c(x) \\ [\bar{c}(x)]^+ \end{bmatrix} \right\|_1$$

(PP):

$$\min_x \; \phi(x,\mu) := \mu f(x) + v(x)$$

(FJ):

$$\mathcal{F}(x,y,\bar{y},\mu) :=$$
$$\mu f(x) + c(x)^T y + \bar{c}(x)^T \bar{y}$$

KKT residual:

$$\rho(x,y,\bar{y},\mu)$$

Local model of ϕ at x_k:

$$l_k(d,\mu)$$

KKT conditions for (FP) and (PP) expressed with residual

$$\rho(x,y,\bar{y},\mu) := \begin{bmatrix} \mu g(x) + J(x)y + \bar{J}(x)\bar{y} \\ \min\{[c(x)]^+, e - y\} \\ \min\{[c(x)]^-, e + y\} \\ \min\{[\bar{c}(x)]^+, e - \bar{y}\} \\ \min\{[\bar{c}(x)]^-, \quad \bar{y}\} \end{bmatrix}$$

KKT conditions for (QP) expressed with

$$\rho_k(d,y,\bar{y},\mu,H) := \begin{bmatrix} \mu g_k + Hd + J_k y + \bar{J}_k \bar{y} \\ \min\{[c_k + J_k^T d]^+, e - y\} \\ \min\{[c_k + J_k^T d]^-, e + y\} \\ \min\{[\bar{c}_k + \bar{J}_k^T d]^+, e - \bar{y}\} \\ \min\{[\bar{c}_k + \bar{J}_k^T d]^-, \quad \bar{y}\} \end{bmatrix}$$

Exact solution of (QP):

$$\rho_k(d,y,\bar{y},\mu,H) = 0$$

Assumptions and well-posedness

Assumption

(1) *The functions f, c, and $\bar{c}$ and their first derivatives are bounded and Lipschitz continuous in an open convex set containing $\{x_k\}$ and $\{x_k + d_k\}$.*

(2) *The QP solver can solve (QP) arbitrarily accurately for any $\mu \geq 0$.*

Theorem (Well-posedness)

One of the following holds for our method, iSQO:

1. *iSQO terminates finitely with a KKT point or infeasible stationary point.*
2. *iSQO generates an infinite sequence of iterates*

$$\left(x_k, \begin{bmatrix} y'_k \\ \bar{y}'_k \end{bmatrix}, \begin{bmatrix} y''_k \\ \bar{y}''_k \end{bmatrix}, \mu_k\right) \text{ where } \begin{bmatrix} y'_k \\ y''_k \end{bmatrix} \in [-e, e], \ \begin{bmatrix} \bar{y}'_k \\ \bar{y}''_k \end{bmatrix} \in [0, e], \text{ and } \mu_k > 0.$$

Global convergence

Theorem (Global convergence)

One of the following holds:

(a) $\mu_k = \underline{\mu}$ *for some* $\underline{\mu} > 0$ *for all large* k *and either every limit point of* $\{x_k\}$ *corresponds to a KKT point or is an infeasible stationary point;*

(b) $\mu_k \to 0$ *and every limit point of* $\{x_k\}$ *is an infeasible stationary point;*

(c) $\mu_k \to 0$, *all limit points of* $\{x_k\}$ *are feasible, and, with*

$$K_\mu := \{k : \mu_{k+1} < \mu_k\},$$

every limit point of $\{x_k\}_{k \in K_\mu}$ *corresponds to an FJ point where the MFCQ fails.*

Corollary

If $\{x_k\}$ *is bounded and every limit point of this sequence is a feasible point at which the MFCQ holds, then* $\mu_k = \underline{\mu}$ *for some* $\underline{\mu} > 0$ *for all large* k *and every limit point of* $\{x_k\}$ *corresponds to a KKT point.*

"Direct" scenario

(NLP):

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

(FP):

$$\min_x v(x) := \left\| \begin{bmatrix} c(x) \\ [\bar{c}(x)]^+ \end{bmatrix} \right\|_1$$

(PP):

$$\min_x \phi(x, \mu) := \mu f(x) + v(x)$$

(FJ):

$$\mathcal{F}(x, y, \bar{y}, \mu) :=$$
$$\mu f(x) + c(x)^T y + \bar{c}(x)^T \bar{y}$$

KKT residuals:

$$\rho(x, y, \bar{y}, \mu)$$
$$\rho_k(d, y, \bar{y}, \mu, H)$$

Local model of ϕ at x_k:

$$l_k(d, \mu)$$

Terminate the QP solver when the solution $(d_k, y_{k+1}, \bar{y}_{k+1})$ of (QP) with $\mu = \mu_k$ satisfies

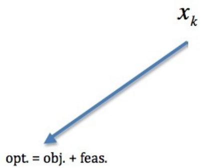
- $y_{k+1} \in [-e, e]$, $\bar{y}_{k+1} \in [0, e]$
- $\Delta l_k(d_k, \mu_k) \geq \theta \|d_k\|^2 > 0$ for $\theta \in (0, 1)$
- $\|\rho_k(d_k, y_{k+1}, \bar{y}_{k+1}, \mu_k, H_k)\| \leq \kappa \|\rho(x_k, y_k, \bar{y}_k, \mu_k)\|$

If

- $\Delta l_k(d_k, \mu_k) \geq \epsilon v_k$ for $\epsilon \in (0, 1)$

then

- $d_k \leftarrow d_k$ is the search direction
- $\mu_{k+1} \leftarrow \mu_k$



"Reference" scenario

(NLP):

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

(FP):

$$\min_x v(x) := \left\| \begin{bmatrix} c(x) \\ [\bar{c}(x)]^+ \end{bmatrix} \right\|_1$$

(PP):

$$\min_x \phi(x, \mu) := \mu f(x) + v(x)$$

(FJ):

$$\mathcal{F}(x, y, \bar{y}, \mu) := \mu f(x) + c(x)^T y + \bar{c}(x)^T \bar{y}$$

KKT residuals:

$$\rho(x, y, \bar{y}, \mu)$$
$$\rho_k(d, y, \bar{y}, \mu, H)$$

Local model of ϕ at x_k:

$$l_k(d, \mu)$$

Terminate the QP solver when the solution $(d_k, y_{k+1}, \bar{y}_{k+1})$ of (QP) with $\mu = 0$ satisfies

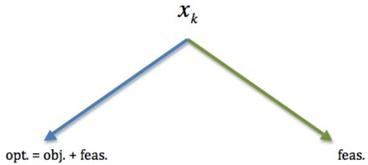
- $y_{k+1} \in [-e, e]$, $\bar{y}_{k+1} \in [0, e]$
- $\Delta l_k(d_k, 0) \ge \theta \|d_k\|^2$ for $\theta \in (0, 1)$
- $\|\rho_k(d_k, y_{k+1}, \bar{y}_{k+1}, 0, H_k)\| \le \kappa \|\rho(x_k, y_k, \bar{y}_k, 0)\|$

If

- $\Delta l_k(d_k, \mu_k) \ge \epsilon \Delta l_k(d_k, 0)$ for $\epsilon \in (0, 1)$

then

- $d_k \leftarrow d_k$ is the search direction
- $\mu_{k+1} \leftarrow \mu_k$



"Combination" scenario

(NLP):

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

(FP):

$$\min_x v(x) := \left\| \begin{bmatrix} c(x) \\ [\bar{c}(x)]^+ \end{bmatrix} \right\|_1$$

(PP):

$$\min_x \phi(x, \mu) := \mu f(x) + v(x)$$

(FJ):

$$\mathcal{F}(x, y, \bar{y}, \mu) :=$$
$$\mu f(x) + c(x)^T y + \bar{c}(x)^T \bar{y}$$

KKT residuals:

$$\rho(x, y, \bar{y}, \mu)$$
$$\rho_k(d, y, \bar{y}, \mu, H)$$

Local model of ϕ at x_k:

$$l_k(d, \mu)$$

Choose the largest $\tau \in [0, 1]$ such that

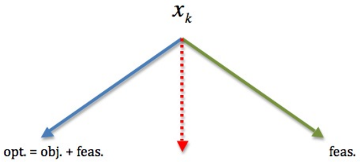
$$d_k \leftarrow \tau d_k + (1 - \tau) d_k$$

yields

$$\Delta l_k(d_k, 0) \ge \epsilon \Delta l_k(d_k, 0)$$

then choose $\mu_{k+1} < \mu_k$ such that

$$\Delta l_k(d_k, \mu_{k+1}) \ge \beta \Delta l_k(d_k, 0) \text{ for } \beta \in (0, 1)$$



iSQO framework

repeat

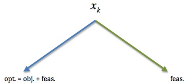
(1) Check whether KKT point or infeasible stationary point has been obtained.

(2) Compute an inexact solution of (QP) with $\mu = \mu_k$.

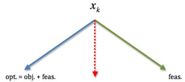
(a) If “Direct” scenario occurs, then go to step 4.

(3) Compute an inexact solution of (QP) with $\mu = 0$.

(a) If “Reference” scenario occurs, then go to step 4.



(b) If “Combination” scenario occurs, then go to step 4.



(4) Perform a backtracking line search to reduce $\phi(\cdot, \mu_{k+1})$.

endrepeat

Complicating factors

A few special cases make our actual algorithm slightly ;-) more complicated

- Landing on stationary points for $\phi(\cdot, \mu_k)$
  - We allow only a multiplier and/or penalty parameter update
- A tightened accuracy tolerance is needed in “combination” scenarios
  - We may require certain multipliers to be close to their bounds
  - (Think of identifying violated constraints)
- H_k and/or H_k may not be positive definite
  - We ask the QP solver to check the curvature along trial directions
  - (Dynamic inertia correction if trial curvature is too small/negative)

Actual algorithm involves six scenarios, but we have presented the “core” ideas

Implementation details: Experiments in paper

- Matlab implementation
- Test set involves 309 `CUTEr` problems with
  - at least one free variable
  - at least one general (non-bound) constraint
  - #variables + #constraints $\leq$ 20,000 (because it's Matlab!)
  - no failures due to `BQPD`
- `BQPD` for QP solves with indefinite Hessians; see (Fletcher, 2000)
- *Simulated* inexactness by perturbing QP solutions
- Termination conditions ($\epsilon_{tol} = 10^{-6}$ and $\epsilon_{\mu} = 10^{-8}$):

$$\|\rho(x_k, y_k, \bar{y}_k, \mu_k)\|_\infty \leq \epsilon_{tol} \quad \text{and} \quad v_k \leq \epsilon_{tol}; \qquad \text{(Optimal)}$$

$$\|\rho(x_k, y_k, \bar{y}_k, 0)\|_\infty = 0 \quad \text{and} \quad v_k > 0; \qquad \text{(Infeasible)}$$

$$\|\rho(x_k, y_k, \bar{y}_k, 0)\|_\infty \leq \epsilon_{tol} \quad \text{and} \quad v_k > \epsilon_{tol} \quad \text{and} \quad \mu_k \leq \epsilon_\mu \qquad \text{(Infeasible)}$$

- Investigate performance of inexact algorithm with $\kappa = 0.01$, 0.1, and 0.5

Success statistics

Counts of termination messages for exact and three variants of inexact algorithm:

Termination message	Exact	Inexact		
		$\kappa = 0.01$	$\kappa = 0.1$	$\kappa = 0.5$
Optimal solution found	289	291	293	293
Infeasible stationary point found	8	7	6	6
Iteration limit reached	12	11	10	10

Termination statistics and reliability do not degrade with inexactness!

Inexactness levels

Observe "induced" relative residuals for QP solves:

$$\kappa_I := \frac{\|\rho_k\|}{\|\rho\|}$$

For problem j, we compute minimum $(\kappa_I(j))$ and mean $(\bar{\kappa}_I(j))$ values over run:

min	κ	$\bar{\kappa}_{I,\min}$	$[0-10^{-8})$	$[10^{-8}-10^{-6})$	$[10^{-6}-10^{-4})$	$[10^{-4}-10^{-3})$	$[10^{-3}-0.01)$	$[0.01-0.1)$	$[0.1-0.5)$	$[0.5-1)$	$[1-\infty)$
$\kappa_I(j)$	0.01	3.8e-03	0	1	9	6	277	1	0	0	0
	0.1	3.0e-02	0	0	1	10	30	254	0	0	0
	0.5	9.0e-02	0	0	2	4	24	77	188	0	0
mean	κ	$\bar{\kappa}_{I,\text{mean}}$									
$\bar{\kappa}_I(j)$	0.01	7.4e-03	0	0	0	0	270	24	0	0	0
	0.1	7.0e-02	0	0	0	0	0	280	15	0	0
	0.5	3.5e-01	0	0	0	0	0	1	282	12	0

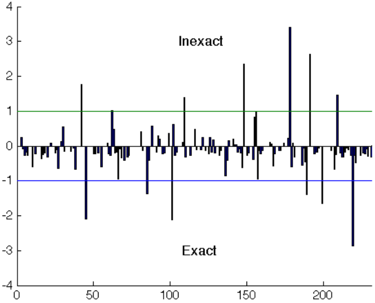
Relative residuals generally need only be moderately smaller than parameter κ!

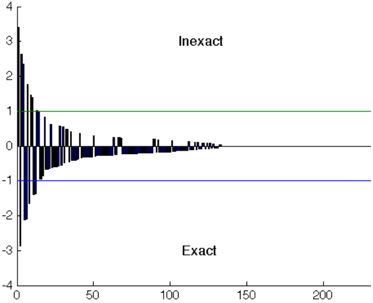
Iteration comparison

Considering the logarithmic outperforming factor

$$r^j := -\log_2(\text{iter}^j_{\text{inexact}}/\text{iter}^j_{\text{exact}}),$$

we compare iteration counts of our inexact ($\kappa = 0.01$) and exact algorithms:





Iteration counts do not degrade significantly with inexactness!

Implementation details: Experiments from last night

- Matlab implementation
- Test set involves 60 **CUTEr** problems that were all successfully solved with
  - augmented Lagrangian (AL) method
  - sequential quadratic optimization (SQO) method
  - iSQO method with AL method for QP subproblems
- AL method or **CPLEX** for QP solves: switch when

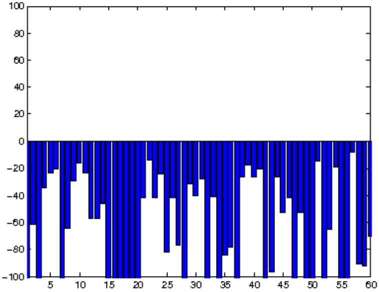
$$\|\rho(x_k, y_k, \bar{y}_k, \mu_k)\|_\infty \leq 10^{-2} \quad \text{and} \quad v_k \leq 10^{-2}$$

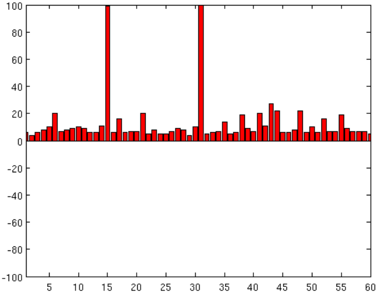
- Terminate solve for (NLP) when

$$\|\rho(x_k, y_k, \bar{y}_k, \mu_k)\|_\infty \leq 10^{-6} \quad \text{and} \quad v_k \leq 10^{-6}$$

Iteration comparison: AL vs. SQO

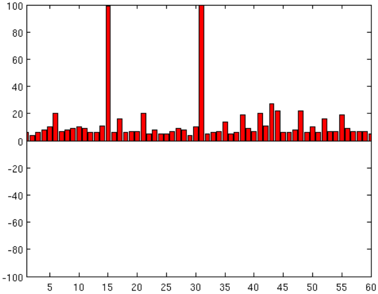
AL (left) with "cheap" iterations vs. SQO (right) with "expensive" iterations

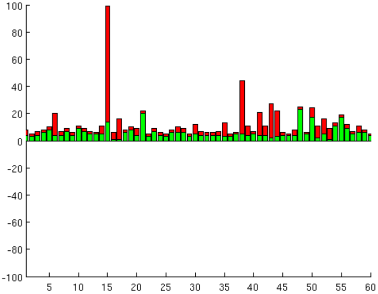




Iteration comparison: SQO vs. iSQO

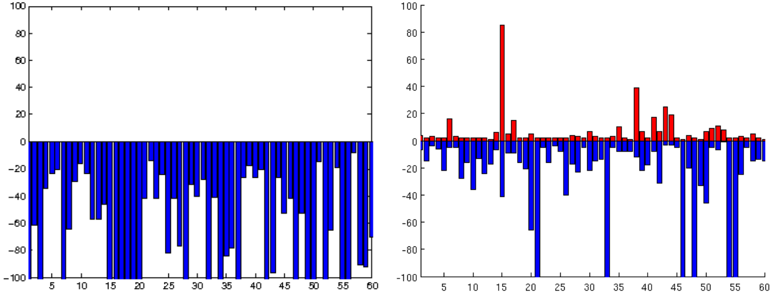
SQO (left) with “expensive” iterations vs. iSQO (right) with “cheaper” iterations





Iteration comparison: AL vs. iSQO(subproblems)

AL (left) with "cheap" iterations vs. iSQO (right) with few "expensive" iterations



Summary

Contributions:

- Developed, analyzed, and experimented with an inexact SQO method
- Allows generic inexactness in QP subproblem solves
- No specific QP solver required
- Global convergence guarantees established
- Numerical experiments suggest inexact algorithm is reliable
- Inexact solutions allowed without degradation of performance

Thanks!

“Exact” Algorithms:

- J. V. Burke, F. E. Curtis, and H. Wang, “A Sequential Quadratic Optimization Algorithm with Rapid Infeasibility Detection,” to appear in *SIAM Journal on Optimization*, 2014.
- R. H. Byrd, F. E. Curtis, and J. Nocedal, “Infeasibility Detection and SQP Methods for Nonlinear Optimization,” *SIAM Journal on Optimization*, Volume 20, Issue 5, pg. 2281-2299, 2010.

“Inexact” Algorithms:

- F. E. Curtis, T. C. Johnson, D. P. Robinson, and A. Wächter, “An Inexact Sequential Quadratic Optimization Algorithm for Large-Scale Nonlinear Optimization,” to appear in *SIAM Journal on Optimization*, 2014.
- F. E. Curtis, J. Huber, O. Schenk, and A. Wächter, “A Note on the Implementation of an Interior-Point Algorithm for Nonlinear Optimization with Inexact Step Computations,” *Mathematical Programming, Series B*, Volume 136, Issue 1, pg. 209–227, 2012.
- F. E. Curtis, O. Schenk, and A. Wächter, “An Interior-Point Algorithm for Large-Scale Nonlinear Optimization with Inexact Step Computations,” *SIAM Journal on Scientific Computing*, Volume 32, Issue 6, pg. 3447-3475, 2010.
- R. H. Byrd, F. E. Curtis, and J. Nocedal, “An Inexact Newton Method for Nonconvex Equality Constrained Optimization,” *Mathematical Programming*, Volume 122, Issue 2, pg. 273-299, 2010.
- F. E. Curtis, J. Nocedal, and A. Wächter, “A Matrix-free Algorithm for Equality Constrained Optimization Problems with Rank-Deficient Jacobians,” *SIAM Journal on Optimization*, Volume 20, Issue 3, pg. 1224-1249, 2009.
- R. H. Byrd, F. E. Curtis, and J. Nocedal, “An Inexact SQP Method for Equality Constrained Optimization,” *SIAM Journal on Optimization*, Volume 19, Issue 1, pg. 351-369, 2008.