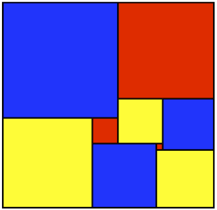
Efficient Solutions for the Large Scale Trust Region Subproblem, TRS

Henry Wolkowicz

(work with Ting Kei Pong and Heng Ye)
Dept. of Combinatorics and Optimization
University of Waterloo



Success of Linear Models/LP/Duality

Classical Linear Program

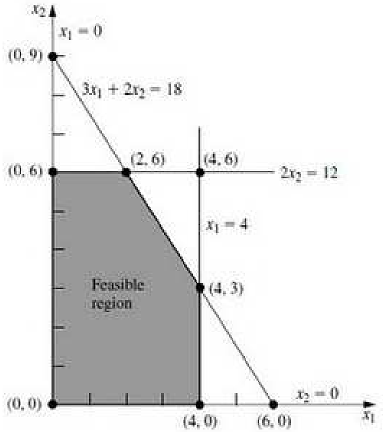
$$\text{(PLP)} \qquad \begin{array}{rll} p^* := & \min & f(y) := b^T y \\ & \text{s.t.} & g(y) := c - A^T y \leq 0 \\ & & x \in \mathbb{R}^n \end{array}$$

Strong Duality if (LP) Feasible; and Attained if Finite

$$\text{(DLP)} \qquad \begin{array}{rcl} p^* & = & d^* \\ & := & \max_{x \geq 0} \min_y L(y,x) := b^T y + x^T (c - A^T y) \\ & := & \max_{x \geq 0} \min_y c^T x + y^T (b - Ax) \\ & := & \max_{x \geq 0} \{c^T x : Ax = b, x \geq 0\} \text{ hidden constraint} \end{array}$$

Duality is behind algorithms (simplex, interior-point).

Polyhedral (so Convex) Feasible Set



Quadratic Models

Modelling NP-Hard Problems

$$\text{(MaxCut)} \qquad p^* := \begin{array}{ll} \max & q_0(x) := x^T L x \\ \text{s.t.} & x_i \in \{\pm 1\}, x \in \mathbb{R}^n \end{array}$$

(L is Laplacian matrix of the graph)
Then, equivalent quadratic constraints to ± 1 are:

$$p^* = \max\{x^T L x : x_i^2 = 1, i = 1, \ldots, n\}$$

Strong Duality Fails; No Lagrange Multiplier

Simple example:

$$\text{(P)} \qquad p^* := \begin{array}{ll} \min & f(x) := x \\ \text{s.t.} & g(x) := x^2 \leq 0 \end{array}$$

$$0 = \nabla L(x^*, \lambda) = \nabla f(x^*) + \lambda \nabla g(x^*) = 1 + \lambda 2(0) \qquad \text{impossible}$$

Quadratic Models; Trust Region Subproblem, (TRS)

"Nonconvex" quadratic minimization

$$\text{(TRS)} \qquad \begin{array}{rll} q^* = & \min & q(x) := x^T A x - 2a^T x \\ & \text{s.t.} & \|x\|^2 \le s^2, \\ & & x \in \mathbb{R}^n \end{array}$$

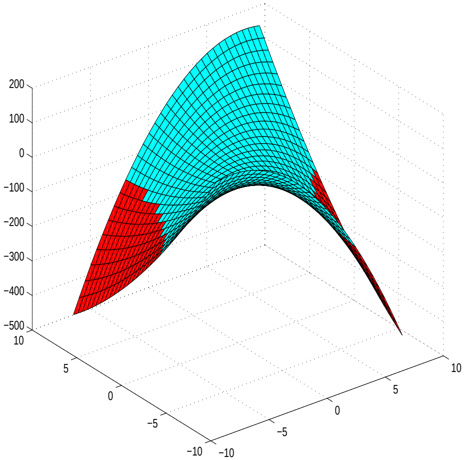
- $A \in \mathcal{S}^n$ - $n \times n$ symmetric (possibly indefinite) matrix
- $a \in \mathbb{R}^n$; $s > 0$ (TR radius);

Generalized TRS, (GTRS)

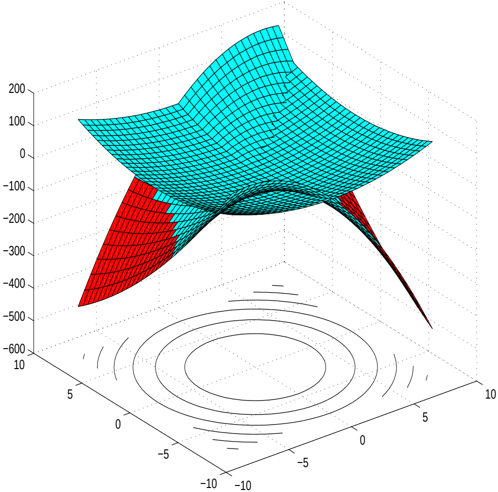
Indefinite Objective; Indefinite Two-sided Quadratic Constraint

$$\text{(GTRS)} \qquad \begin{array}{rll} q^* = & \min & q(x) := x^T A x - 2a^T x \\ & \text{s.t.} & \ell \le q_1(x) := x^T B x - 2b^T x \le u \end{array}$$

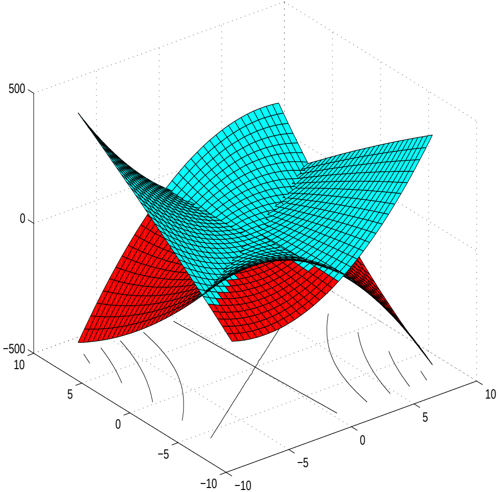
Indefinite Quadratic (Surface Plot, Saddle)



Indefinite Quadr. with Euclid. Norm Constr., TRS



Indefinite Quadr. with Indefinite Quadr. Constr., GTRS



Why Convexity Matters

R. Tyrrell Rockafellar, in SIAM Review, 1993 [9]

"...in fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity."

If Strong Duality Holds; We Get Implicit Convexity

if

$$\begin{aligned} p^* = d^* := \quad & \max_{\lambda \in \Lambda} \min_{x \in \Omega} & L(x, \lambda) := f(x) + \lambda^T g(x) \\ & = \max_{\lambda \in \Lambda} & \phi(\lambda) \end{aligned}$$

$\phi(\lambda)$ is concave (a minimum of functions linear in λ); therefore the dual is: concave maximization

(However, the evaluation of $\phi(\lambda)$ may be costly.)

Outline

Background

- Generalized Trust Region Subproblem and Characterizations of Optimality
- MoSo (1983) algorithm for TRS; within a primal-dual framework; handling the hard case

Duality/Exploiting Sparsity & Degeneracy/Hard Case

- Various primal-dual pairs with strong duality
- Parametric eigenvalue problem and sparsity
- Hard case is actually easiest case, with an explicit solution
- Underdetermined least squares, Numerics

Applications

Many important applications:

- forming subproblems for constrained optimization
- REGULARIZATION of ill-posed problems (inverse image denoising, Electrical impedance tomography, EIT, MRI, NMRI, MRT, CT PET...)
- theoretical applications
- **trust region (TR) methods for unconstr. min.** uses TRS formed from second order Taylor series approximation

Characterization of Optimality for TRS

Surprising: TRS (GTRS) is *on watershed* of convex/nonconvex

- a characterization of optimality of a (possibly) nonconvex problem
- second order positive semidefinite necessary conditions hold on all of $\mathbb{R}^n$

Characterization of x^* optimal for TRS (iff)

(Gay-81 [2], More-Sorensen-83 [5])

$$\left.\begin{array}{c} (A-\lambda^* I)x^* = a, \\ A-\lambda^* I \succeq 0, \lambda^* \leq 0 \end{array}\right\} \quad \text{dual feasibility}$$

$$\|x^*\|^2 \leq s^2 \quad \text{primal feasibility}$$

$$\lambda^*(s^2 - \|x^*\|) = 0 \quad \text{complementary slackness}$$

Characterization of Optimality for (homog) GTRS

(Mild) Slater Type Constraint Qualification

$$\text{(GTRS)} \qquad q^* = \begin{array}{ll} \min & q(x) := x^T A x - 2a^T x \\ \text{s.t.} & \ell \leq q_1(x) := x^T B x \leq u \end{array}$$

$$(CQ) \qquad \exists \hat{x} \text{ s.t. } l < q_1(\hat{x}) := \hat{x}^T B \hat{x} < u$$

Characterization of x^* optimal for GTRS (iff)

(More-93 [4], Stern-W-93 [10])

$$\left.\begin{array}{c} (A - \lambda^* B) x^* = a, \\ A - \lambda^* B \succeq 0 \end{array}\right\} \qquad \text{dual feasibility}$$

$$\ell \leq {x^*}^T B x \leq u \qquad \text{primal feasibility}$$

$$\lambda^*(u - {x^*}^T B x^*) \geq 0 \geq \lambda^*({x^*}^T B x^* - \ell) \qquad \text{compl. slack.}$$

S-Lemma Yakubovich-71,73

(survey in Polik-Terlaky:07 [6] (Farkas Lemma for quadratic functions))

$f, g : \mathbb{R}^n \to \mathbb{R}$ quadratic functions
$\exists \bar{x}$ s.t. $g(\bar{x}) < 0$
then TFAE:

1. $\nexists x \in \mathbb{R}^n : f(x) < 0, g(x) \leq 0$
2. $\exists \lambda \geq 0$ s.t. $f(x) + \lambda g(x) \geq 0, \forall x \in \mathbb{R}^n$

Finsler:37 , Hestenes-McShane:40
$\{x^T Bx = 0, x \neq 0 \implies x^T Ax > 0\}$
implies
$\{\exists \lambda \geq 0, A + \lambda B \succ 0\}$

Mo-So Algorithm Framework

Recall: Opt. conditions of TRS

$$\left.\begin{array}{c}(A-\lambda^* I)x^* = a,\\ A-\lambda^* I \succeq 0, \lambda^* \leq 0\end{array}\right\}$$ dual feasibility

$$\|x^*\|^2 \leq s^2$$ primal feasibility

$$\lambda^*(s^2 - \|x^*\|) = 0$$ complementary slackness

Original view: if $A - \lambda^* I \succ 0, \lambda^* < 0$

- define: $x(\lambda) = (A-\lambda I)^{-1} a$
- <u>SOLVE</u>: $\psi(\lambda) := \|x(\lambda)\| - s = \|(A-\lambda I)^{-1} a\| - s = 0$
- maintain: $A - \lambda I \succ 0, \ \lambda \leq 0$

Less Nonlinear $\phi(\lambda)$ for Newton's Method

Newton for $\psi(\lambda) = 0$?

$A = Q\Lambda Q^T$ orthogonal diagonalization; $\gamma = Q^T a$;

$\psi(\lambda) = \|x(\lambda)\|^2 - s^2 = \sum_{j=1}^{n} \frac{\gamma_j^2}{(\lambda_j(A)-\lambda)^2} - s^2$ *highly* nonlinear in λ, in particular near $\lambda_1(A)$.

less nonlinear $\phi(\lambda)$ (Reinsch:67 [7, 8], Hebden:73 [3])

SOLVE: $\phi(\lambda) := \frac{1}{s} - \frac{1}{\|x(\lambda)\|} = 0$

Newton iterates $\lambda^{(k)}$ for $\phi(\lambda) = 0$; exploit Cholesky

$$\begin{aligned} \lambda^{(k+1)} &= \lambda^{(k)} \boxed{-\frac{\phi(\lambda^{(k)})}{\phi'(\lambda^{(k)})} \text{ Newton step}} \\ &= \lambda^{(k)} - \left(\frac{\|x\|}{\|y\|}\right)^2 \left(\frac{\|x\|-s}{s}\right), \end{aligned}$$

where: $A - \lambda^{(k)} I = R^T R$ (Cholesky); $R^T R x = a$; $R^T y = x$

Computing Newton direction/step Gay:81, MoSo:83

Assume $\lambda^{(k)} \leq 0$ and $A - \lambda^{(k)} I \succ 0$ (i.e. $\lambda^{(k)} < \lambda_1(A)$)

- Factor $A - \lambda^{(k)} I = R^T R$ (Cholesky factorization).
- Solve for x, $R^T R x = a$ $\quad$ ($x = x(\lambda^{(k)})$).
- Solve for y, $R^T y = x$.
- Let $\lambda^{(k+1)} = \lambda^{(k)} - \left[\frac{\|x\|}{\|y\|}\right]^2 \left[\frac{(\|x\| - s)}{s}\right]$ $\quad$ (Newton step); but, *safeguard/maintain* positive definiteness $A - \lambda^{(k+1)} I \succ 0$

Nearly linear

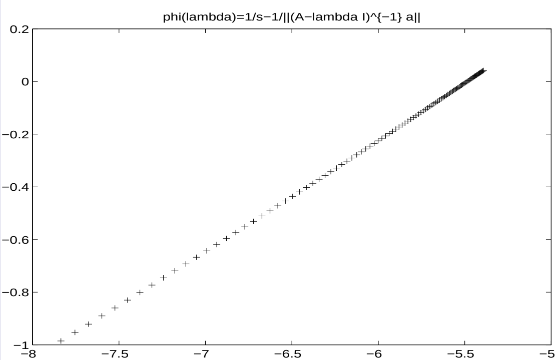


Figure: "Easy" case: $\phi(\lambda) = \frac{1}{s} - \frac{1}{||(A-\lambda I)^{\dagger}a||}$

“Easy” Case; Newton Quadr. Cvgnce from One Side

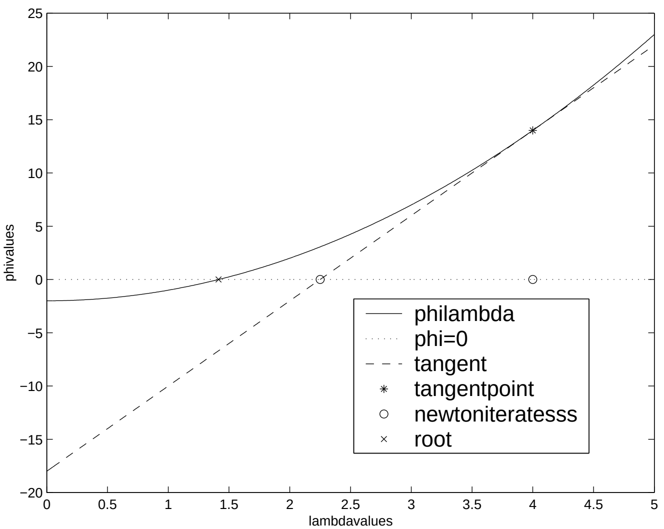


Figure: Newton’s method with the secular function, $\phi(\lambda)$.

Easy/Hard Cases for TRS; Singularity of Hessian

$\lambda_1(A) = \lambda_{\min}(A) > \lambda^* \implies A - \lambda^* I \succ 0$

1. Easy case	2.(a) Hard case (case 1)	2.(b) Hard case (case 2)
$a \notin \mathcal{R}(A - \lambda_1(A)I)$ (stationarity implies $\lambda^* \neq \lambda_1(A)$, so that $\lambda^* < \lambda_1(A)$, Hessian is pos. def., nonsingular)	$a \perp \mathcal{N}(A - \lambda_1(A)I)$ but $\lambda^* < \lambda_1(A)$ (Hessian is pos. def., nonsingular)	$a \perp \mathcal{N}(A - \lambda_1(A)I)$ and $\lambda^* = \lambda_1(A)$ singular Hessian (i) $\\|(A - \lambda^* I)^\dagger a\\| = s$ or $\lambda^* = 0$ (ii) $\\|(A - \lambda^* I)^\dagger a\\| < s$, $\lambda^* < 0$

three different cases for the trust region subproblem;
two subcases (i) and (ii) for the hard case (case 2), where
$A - \lambda^* I$ is positive semidefinite, SINGULAR.
($\cdot^\dagger$ denotes Moore-Penrose gen. inverse.)

Hard Case is really easiest?

We take advantage of Lanczos algorithm in large sparse case.
Will see: hard case becomes easiest case.

$\phi(\lambda)$ in the (near) Hard Case

Newton steps (linearizations) are *"useless"*

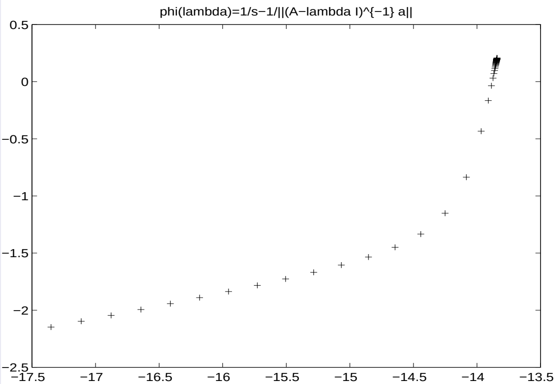


Figure: (near) hard case case: $\phi(\lambda) = \frac{1}{s} - \frac{1}{||(A-\lambda I)^{\dagger} a||}$

Towards a Parametrization for Large Sparse Case

Motivation: try to avoid Cholesky in large sparse case

Homogenization using $y_0^2 = 1$; optimal value $\mu^* =$

$$= \min_{||x||=s,\ y_0^2=1} x^tAx - 2y_0a^tx \quad \text{homogenize obj.}$$

$$= \max_t \min_{||x||=s,\ y_0^2=1} x^tAx - 2y_0a^tx + ty_0^2 - t$$

$$\geq \max_t \min_{||x||^2+y_0^2=s^2+1} x^tAx - 2y_0a^tx + ty_0^2 - t \quad \boxed{\text{**eig prob**}}$$

$$\geq \max_{t,\lambda} \min_{x,y_0} x^tAx - 2y_0a^tx + ty_0^2 - t + \lambda(||x||^2 + y_0^2 - s^2 - 1)$$

$$= \max_{r,\lambda} \min_{x,y_0} x^tAx - 2y_0a^tx + ry_0^2 - r + \lambda(||x||^2 - s^2)$$

$$= \max_\lambda \left(\max_r \min_{x,y_0} x^tAx - 2y_0a^tx + ry_0^2 - r + \lambda(||x||^2 - s^2) \right)$$

$$= \max_\lambda \min_{x,y_0^2=1} x^tAx - 2y_0a^tx + \lambda(||x||^2 - s^2)$$

$$= \mu^*,$$

equated r with $t + \lambda$.

used strong duality for TRS for last two equalities.

Parametric Eigenvalue Problem

$$\max_t \boxed{\min_{||x||^2+y_0^2=s^2+1} x^tAx - 2y_0a^tx + ty_0^2} - t$$

$$k(t) := (s^2+1)\lambda_1(D(t)) - t \qquad D(t) = \begin{bmatrix} t & -a^t \\ -a & A \end{bmatrix}$$

An Unconstrained dual problem to (TRS)

$$\max_t k(t) \qquad \text{(concave max)}$$

$k'(t) = (s^2 + 1)y_0(t)^2 - 1$

If $\lambda_1(D(t))$ is simple

- $y(t)$ normalized eigenvector for $\lambda_1(D(t))$
- $y(t) = \begin{pmatrix} y_0(t) \\ x(t) \end{pmatrix}$
- $$\frac{1}{y_0(t)}\|x(t)\| = s \text{ if, and only if, } k'(t) = 0$$

Linear Primal-Dual SDP Pair

From parametric eigenvalue problem

$$(DSDP) \qquad \begin{array}{rll} \mu^* = & \max_{\lambda,t} & (s^2+1)\lambda - t \\ & \text{s.t.} & D(t) \succeq \lambda I \qquad (\lambda_{\min}(D(t)) \geq \lambda) \end{array}$$

(with $\lambda \leq 0$ for inequality constraint TRS)

$$(PSDP) \qquad \begin{array}{rll} \mu^* = & \min & \operatorname{trace} D(0)X \\ & \text{s.t.} & \operatorname{trace} X = s^2+1 \\ & & X_{11} = 1 \\ & & X \succeq 0 \end{array}$$

$$X \cong \begin{pmatrix} 1 \\ x \end{pmatrix} \begin{pmatrix} 1 \\ x \end{pmatrix}^T$$

Hard Case is Easiest Case: Shift and Deflation

If $\lambda_{\min} < 0$ use $A \leftarrow A - \lambda_{\min}(A)$

- hard case (case 2) happens only when $\lambda^* = \lambda_{\min}(A) < 0$
- $A \succeq 0$ after the shift (objective changes by a constant)

Lemma Fortin-W:03 [1]

$A = \sum_{i=1}^{n} \lambda_i(A) v_i v_i^T = Q \Lambda Q^T$
orthogonal spectral decomposition of A; with $\gamma_i = (Q^T a)_i$

$$\begin{aligned} S_1 &= \{i : \gamma_i \neq 0, \lambda_i(A) > \lambda_{\min}(A)\} \\ S_2 &= \{i : \gamma_i = 0, \lambda_i(A) > \lambda_{\min}(A)\} \\ S_3 &= \{i : \gamma_i \neq 0, \lambda_i(A) = \lambda_{\min}(A)\} \\ S_4 &= \{i : \gamma_i = 0, \lambda_i(A) = \lambda_{\min}(A)\} \end{aligned}$$

For $k = 1, 2, 3, 4,$: $A_k = \sum_{i \in S_k} \lambda_i(A) v_i v_i^T$
Then:

Lemma Conclusions

1. If $S_3 \neq \emptyset$ (easy case), then (x^*, λ^*) solves TRS iff (x^*, λ^*) solves TRS when A is replaced by $A_1 + A_3$.
2. If $S_3 = \emptyset$ (hard case), let $i_0 = 1 \in S_4$, then
   (x^*, λ^*) solves TRS iff
   (x^*, λ^*) solves TRS when A is replaced by
   $\boxed{A_1 + \lambda_{i_0}(A) v_{i_0} v_{i_0}^T.}$ (DEFLATE)
3. Let $x(\lambda^*) = (A - \lambda^* I)^\dagger a$, then (x^*, λ^*), where $x^* = x(\lambda^*) + z, z \in \mathcal{N}(A - \lambda^* I)$ and $\|x^*\| = s$ solves TRS iff $(x(\lambda^*), \lambda^* - \lambda_{\min}(A))$ solves TRS when A is replaced by $A - \lambda_{\min}(A) I$.
4. If $\lambda_{\min}(A) \geq 0$, then (x^*, λ^*) solves TRS iff
   (x^*, λ^*) solves TRS when A is replaced by
   $A + \sum_{i \in S_4} \alpha_i v_i v_i^T$, with $\alpha_i \geq 0$. (DEFLATE all in S_4)

Shift/Deflate: Solve Hard Case Explicitly

Summary - sometimes you get !lucky!

- eigenpair $\lambda_{\min}(A) = \lambda_1(A), v_1$ found; large sparse case; so, assume possible hardcase: $\boxed{\lambda_1(A) < 0 \text{ and } v_1^T a = 0}$
- !we have $\lambda_1(A)$! so shift: $A \leftarrow A - \lambda_{\min}(A) I \succeq 0$; !optimum unchanged, objective value changed by $\lambda_1 s^2$!.
- !we have v_1, $v_1^T a = 0$! so deflate: $A \leftarrow A + \alpha_1 v v^T$, $\|A\| > \alpha_1 >> 0$; (repeat deflation if needed, i.e as long as $\boxed{\lambda_{\min}(A) = 0 \text{ and } v^T a = 0}$) !objective value/optima unchanged!.
- if $v^T a \neq 0$, then continue with TRS algorithm, i.e. hard case (case 1); otherwise, $A \succ 0$, so (using prec. conj grad.) calculate $\bar{x} = A^{-1} a$; if $\|x(\lambda^*)\| > s$, then continue with TRS algorithm, i.e. hard case (case 1); otherwise $\|x(\lambda^*)\| \leq s$, then we have hard case (case 2) and !we have an explicit solution : $\|x^*\| = \|\bar{x} + \beta v_1\| = s$, !since $v_1 \in \mathcal{N}(A - \lambda^* I)$!.

cputime in the hard case (*case 2*)

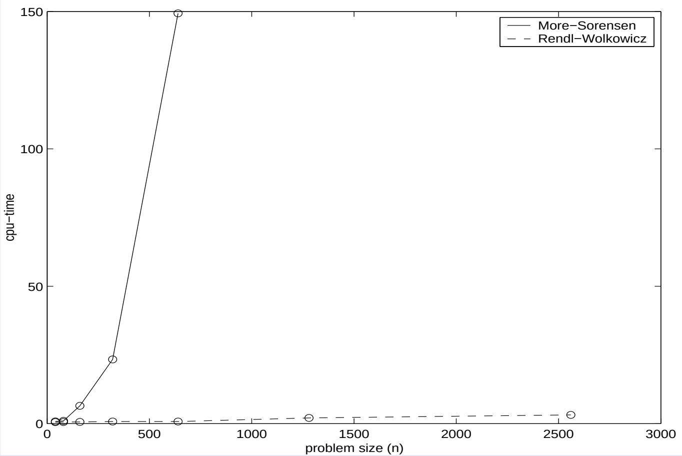


Figure: cputime in the hard case (*case 2*)

log of cputime in the hard case (*case 2*)

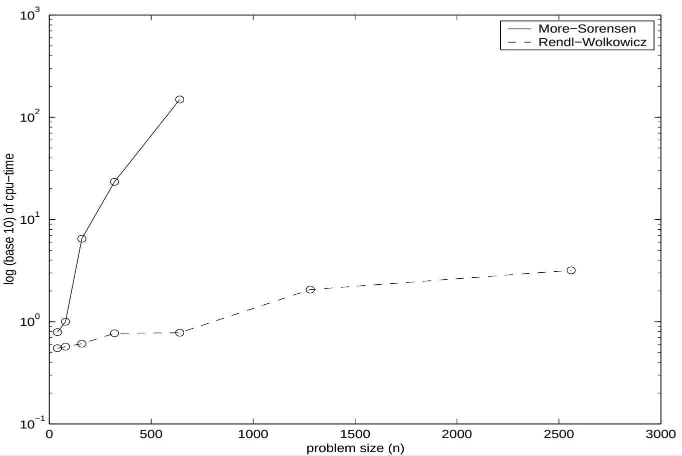


Figure: log of cputime in the hard case (*case 2*)

Conclusion

Importance of Duality/Convexity; Exploit Singularity

- The TRS (quadratic objective, one quadratic constraint) is an *implicit* convex problem.
- Convexity allows one to exploit the special structure of the dual along with information from the primal problem.
- Singularity is usually an indication of difficulty/ill-posedness. However, if the structure is well known, then one can often exploit singularity to improve the algorithm.

- C. Fortin and H. Wolkowicz, *The trust region subproblem and semidefinite programming*, Optim. Methods Softw. **19** (2004), no. 1, 41–67. MR MR2062235 (2005c:90065)
- D.M. Gay, *Computing optimal locally constrained steps*, SIAM J. Sci. Statist. Comput. **2** (1981), 186–197.
- M.D. Hebden, *An algorithm for minimization using exact second derivatives*, Tech. Report TP515, Atomic Energy Research Establishment, Harwell, England, 1973.
- J.J. Moré, *Generalizations of the trust region problem*, Optim. Methods Software **2** (1993), 189–209.
- J.J. Moré and D.C. Sorensen, *Computing a trust region step*, SIAM J. Sci. Statist. Comput. **4** (1983), 553–572.
- I. Pólik and T. Terlaky, *A survey of the S-lemma*, SIAM Rev. **49** (2007), no. 3, 371–418 (electronic). MR MR2353804
- C. Reinsch, *Smoothing by spline functions*, Numer. Math. **10** (1967), 177–183.

______, *Smoothing by spline functions II*, Numer. Math. **16** (1971), 451–454.

R. Tyrrell Rockafellar, *Lagrange multipliers and optimality*, SIAM Rev. **35** (1993), no. 2, 183–238. MR 1220880 (94h:49004)

R. Stern and H. Wolkowicz, *Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations*, SIAM J. Optim. **5** (1995), no. 2, 286–313. MR 96h:90077

Thanks for your attention!

Efficient Solutions for the Large Scale Trust Region Subproblem, TRS

Henry Wolkowicz

(work with Ting Kei Pong and Heng Ye)
Dept. of Combinatorics and Optimization
University of Waterloo