A Survey of the Trust Region Subproblem

Henry Wolkowicz

Dept. Comb. and Opt., University of Waterloo, Canada

Thur. Feb. 9, 10:00-10:20 (EST) 2023

Outline

- Background, Duality
- MS (1983) and GLTR (1999) algorithms within SDP Framework
- RW (1997) and FW (2002) Algorithm (Revisited)
- Robustness, Exploiting Sparsity
- Numerics
- Emphasis: Hard Case - really hard?

$$\text{(UNC)} \qquad \mu^* := \min_{x \in \mathbb{R}^n} f(x)$$

Quadratic Model at current estimate x_c:

$$\text{(Quad)} \qquad \begin{array}{ll} \min & f(x_c) + \nabla f(x_c)^T d + \frac{1}{2} d^T \nabla^2 f(x_c) d \\ \text{s.t.} & \|d\| \le s. \end{array}$$

The optimal d exists and can be found efficiently.

The Trust Region Subproblem

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

A, $n \times n$ symmetric (possibly indefinite) matrix
a, n-vector; $s > 0$, TR radius
q is (possibly) nonconvex quadratic

Many Applications

- subproblems for constrained optimization
- regularization of ill-posed problems
- theoretical applications
- etc...
- **trust region (TR) methods**

Many Advantages for TR, e.g.:
second order optimality conditions
q-quadratic convergence
BUT: popularity? sparsity? hard case?

Special Case: LLS/Regularization

- find approx. solutions for LLS

$$\text{LLS} \quad \min_x \|Gx - d\|^2, \quad G \text{ singular or ill-cond.}$$

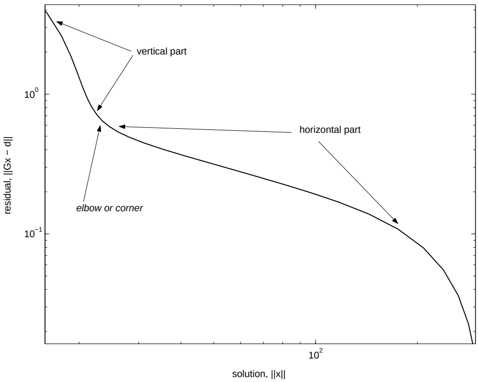
- can be reformulated as a TRS, if an appropriate/correct TR radius $\bar{s}$ can be found

- steps of an efficient TRS algorithm try to find an optimal solution of TRS, $x(\hat{s})$, for a corresponding TR radius $\|x(\hat{s})\| \leq \hat{s}$

L-Curve

From TRS: Find the point of max curvature/elbow

$\mathcal{L}(G, d) = \{(\log(s), \log \|Gx(s) - d\|) : s > 0\}$



$$\text{(TRS)} \qquad \mu(A, a, s) := \begin{array}{ll} \min & q(x) := x^T A x - 2a^T x \\ \text{s.t.} & \|x\|^2 \leq s^2, \end{array}$$

where:

$A = G^T G$, $n \times n$ symmetric (ill-cond.)

$a = G^T d \in \mathbb{R}^n$, $s > 0$, $x \in \mathbb{R}^n$

define λ^* optimal Lagrange multiplier

$x(0) = A^{-1} a = G^{-1} d$ unconstr min

(connection to hard case in TRS)

Optim. Cond.: (Gale-81/Sorensen-82)

x^* optimal for TRS

if and only if

$$\left\{\begin{array}{c} (A-\lambda^* I)x^* = a, \\ \boxed{A-\lambda^* I \succeq 0}, \lambda^* \le 0 \\ \|x^*\|^2 \le s^2 \\ \lambda^*(s^2 - \|x^*\|^2) = 0 \end{array}\right.$$

dual feas. (first two conditions)

primal feas.

compl. slack.

(phrased in the modern primal-dual paradigm) **Surprising:** characterization of opt.; 2nd order psd

Duality and MS Algorithm

For simplicity, use equality TRS:

$$(\text{TRS}_=) \qquad \begin{array}{rll} q^* = & \min & q(x) \\ & \text{s.t.} & \|x\|^2 = s^2, \end{array}$$

Let $\lambda \in \Re$; a lower bound is

$$\min_x q(x) + \lambda(s^2 - \|x\|^2)$$

Hidden constraint

best lower bound yields dual (SDP)

$$\begin{array}{rcl} q^* = \nu^* & := & \max_\lambda \boxed{\min_x x^T(A-\lambda I)x - 2a^Tx + \lambda s^2} \\ & = & \max_\lambda\ h(\lambda) \quad \text{(dual functional)} \end{array}$$

where:
(convex) Lagrangian is $L(x,\lambda) := x^T(A-\lambda I)x - 2a^Tx + \lambda s^2$;
(concave) dual functional is $h(\lambda) := \min_x\ L(x,\lambda)$
$h(\lambda) := \lambda s^2 - a^T(A-\lambda I)^{-1}a$, if $A - \lambda I \succ 0$

Convexity/SDP Arise Naturally

TRS: simpler (constrained) root finding problem

$$h'(\lambda) = 0 \text{ s.t. } \nabla^2 L(x,\lambda) = A - \lambda I \succ 0$$

with $x(\lambda) = (A - \lambda I)^{-1} a$ (Lagrangian stationarity)

---

to maximize concave function h solve:

$$0 = h'(\lambda) = s^2 - x(\lambda)^T x(\lambda)$$

i.e. dual algor: maintain dual feasibility while trying to attain primal feasibility

The Hard Case; $A - \lambda^* I$ Singular

1. Easy case	2.(a) Hard case (case 1)	2.(b) Hard case (case 2)
$a \notin \mathcal{R}(A - \lambda_1(A)I)$ (implies $\lambda^* < \lambda_1(A)$)	$a \perp \mathcal{N}(A - \lambda_1(A)I)$ and $\lambda^* < \lambda_1(A)$	$a \perp \mathcal{N}(A - \lambda_1(A)I)$ and $\lambda^* = \lambda_1(A)$ (i) $\|(A - \lambda^* I)^\dagger a\| = s$ or $\lambda^* = 0$ (ii) $\|(A - \lambda^* I)^\dagger a\| < s, \lambda^* < 0$

Table: 3 cases for TRS; 2 subcases (i), (ii) for hard case (case 2).

$\mathcal{R}(A - \lambda_1(A)I)^\perp = \mathcal{N}(A - \lambda_1(A)I)$ (dimension?)

Handling the Hard Case

LEMMA: spectral decomposition of A:
$A = \sum_{i=1}^{n} \lambda_i(A) v_i v_i^T = P \Lambda P^T$, $P^T P = I$, $\bar{a} = P^T a$.
$A_i := \sum_{i \in S_i} \lambda_i(A) v_i v_i^T, \; i = 1, \ldots, 4$,

$$
\begin{aligned}
S_1 &= \{i : \bar{a}_i \neq 0, \; \lambda_i(A) > \lambda_1(A)\} \\
S_2 &= \{i : \bar{a}_i = 0, \; \lambda_i(A) > \lambda_1(A)\} \\
S_3 &= \{i : \bar{a}_i \neq 0, \; \lambda_i(A) = \lambda_1(A)\} \\
S_4 &= \{i : \bar{a}_i = 0, \lambda_i(A) = \lambda_1(A)\}
\end{aligned}
$$

Define projections: $P_i := \sum_{i \in S_i} v_i v_i^T, \; i = 1, \ldots, 4$.
THEN: $(\bar{a}_i = 0 \Leftrightarrow a^T v_i = 0)$

Shift/deflate 1

Suppose $S_3 \neq \emptyset$ (easy case), $\alpha > 0$, and $i \in S_2 \cup S_4$. Then

$$(x^*, \lambda^*) \text{ solves TRS}$$

<u>**iff**</u>

$$(x^*, \lambda^*) \text{ solves TRS when } A \leftarrow A + \alpha v_i v_i^T.$$

• Equivalently, in the easy case with $a^T v_i = 0$, one can shift (increase) the eigenvalue corresponding to v_i.

Hard-Case Shift

Let $u^* = (A - \lambda^* I)^\dagger a$ with $\|u^*\| < s$ and suppose that $i \in S_2 \cup S_4$ and $\alpha > 0$. Then

$$(x^* = u^* + z, \lambda^*), z \in \mathcal{N}(A - \lambda^* I) \text{ solves TRS}$$

iff

$$(x^* = u^* + z, \lambda^*), z \in \mathcal{N}(A + \alpha v_i v_i^T - \lambda^* I)$$

solves TRS when A is replaced by $A + \alpha v_i v_i^T$.

- Equivalently, in the hard case with $a^T v_i = 0$, one can shift (increase) the eigenvalue corresponding to v_i.

All Case Shift

Let $\lambda_1(A) < 0$; $u^* = (A - \lambda^* I)^\dagger a$. Then

there exists $z \in \mathcal{N}(A - \lambda^* I)$ such that:
$(x^*, \lambda^*),$ with $x^* = u^* + z,$ solves TRS
iff
$(u^*, \lambda^* - \lambda_1(A))$ solves TRS when
$A \leftarrow A - \lambda_1(A) I$

• Equivalently, ensure that $A \succeq 0$ by shifting (increasing) the smallest eigenvalue.
• Implicitly Convex, The equivalent problem is convex, i.e., this shows that TRS is an implicit convex problem; strong duality holds.

Hard Case (case 2); is it ill-posed?

Using shift, wolg assume $A \succeq 0, \lambda^* = 0$.
Then, the difficulty reduces to a regularization question (LLS).

$x^* = A^\dagger a, \ \|x^*\| \leq s$ implies optimum.
$x^* = A^\dagger a, \ \|x^*\| > s$ implies not hard case 2.

Conclusion: Applying regularization might be helpful in determining between the two cases.

Shift with Near-Hard Case

LEMMA: Suppose that x^* solves TRS and $\|x^*\| = s$. Let $\epsilon > 0$ and $v \in \mathcal{R}^n$ with $\|v\| = 1$. Let $\mu^*(\epsilon)$ be the optimal value of TRS when a is perturbed to $a + \epsilon v$. Then

$$-2s\epsilon \leq \mu^* - \mu^*(\epsilon) \leq 2s\epsilon.$$

$\square$

The More-Sorensen (dual) Algor., 1983

Outline:
(i) Use safeguarding/updating: reduce interval of uncertainty $[\lambda_L, \lambda_U]$ for λ^* and improve lower bound λ_S for $\lambda_1(A)$

(ii) Take a Newton step to implicitly solve for λ in $\|(A - \lambda I)^{-1} a\| = s$

(iii) If possible hard case (case 2) is detected (i.e. $\|x(\lambda)\| < s$), then take a *primal step/negative curvature* to the boundary while simultaneously reducing the objective function.

Solving for λ^*

Use orthogonal diagonalization of A; the function $\psi(\lambda)$

$$\|x(\lambda)\| - s = \|Q(\Lambda - \lambda I)^{-1} Q^T a\| - s \approx \frac{c_1}{\lambda_1(A) - \lambda} + d,$$

For some constants $c_1 > 0, d$.

Highly nonlinear for λ near $\lambda_1(A)$ (slow cvgnce)

secular equation

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

(Reinsch and Hebden)
rational structure shows function is less nonlinear

$$\phi(\lambda) \approx \frac{1}{s} - \frac{\lambda_1(A) - \lambda}{c_2},$$

for some $c_2 > 0$.
Also $\phi(\lambda)$ is convex, strictly increasing on $(-\infty, \lambda_1(A))$.

Compute Newton Direction; Take Newton Step

BEGIN algorithm Assume $\lambda_k \leq 0$ and $A - \lambda_k I \succ 0$ (i.e. $\lambda_k < \lambda_1(A)$).

1. Factor $A - \lambda_k I = R^T R$ (Cholesky factorization).
2. Solve, for x, $R^T R x = a$ ($x = x(\lambda_k)$).
3. Solve, for y, $R^T y = x$.
4. Let $\lambda_{k+1} = \lambda_k - \left[\frac{\|x\|}{\|y\|}\right]^2 \left[\frac{(\|x\| - s)}{s}\right]$ (Newton step).

END algorithm

Newton Method on $\phi(\lambda)$

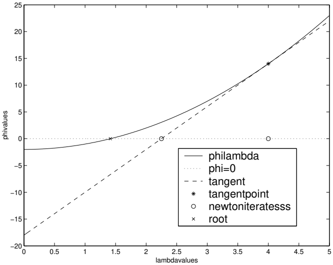


Figure: Newton's method with the secular function, $\phi(\lambda)$.

MS: Handling the Hard Case

LEMMA:[Negative curvature primal step to boundary]
$0 < \sigma < 1$ given; and $A - \lambda I = R^T R$, $(A - \lambda I)x = a$, $\lambda \leq 0$;
$\|x + z\|^2 = s^2$; $\|Rz\|^2 \leq \sigma(\|Rx\|^2 - \lambda s^2)$.
Then

$$|q(x + z) - q(x^*)| \leq \sigma |q(x^*)|.$$

where x^* is optimal for TRS. □

Aside: GLTR Algorithm, 1999

Generalized Lanczos Trust Region (Gould, Lucidi, Roma, Toint)
FOCUS: exploiting sparsity; Lanczos tridiagonalization of A;
solve a *sequence* of restricted problems

$$
(TRS_{\text{sub}}) \quad \begin{array}{ll} \min & q(x) \\ \text{s.t.} & \|x\| \le s \\ & x \in S, \end{array} \quad \text{Krylov subspace}
$$

Outline

(i) Apply CG to $q(\cdot)$ until boundary (or min) is reached; or until direction of nonpositive curvature is found to move to boundary.

(ii) (efficiently) Solve TRS subproblem with constraint $x \in S \equiv \mathcal{K}_k$ where $\mathcal{K}_k := \text{span}\{a, Aa, A^2a, A^3a, \ldots, A^ka\}$

(iii) Increase k, size of Krylov subspace using CG.

For simplicity, use equality TRS:

$$(\text{TRS}_=) \qquad q^* = \begin{array}{ll} \min & q(x) \\ \text{s.t.} & \|x\|^2 = s^2 \end{array}$$

Strong Duality Holds for TRS$_=$

$$q^* = \nu^* := \max_{\lambda} \min_{x} L(x,\lambda),$$ strong duality

$L(x,\lambda) := x^T(A - \lambda I)x - 2a^T x + \lambda s^2,$ Lagrangian
$h(\lambda) := \lambda s^2 - a^T(A - \lambda I)^\dagger a,$ dual functional
THEOREM: (Wolfe Dual)

$$(D) \qquad q^* = \sup_{A - \lambda I \succ 0} h(\lambda)$$

($\succeq$ in easy case, or after shift to $A \succeq 0$) $\square$

Unconstrained and Linear Duals

Homogenizing $\text{TRS}_=$

$$
\begin{array}{rll}
q^* = & \min & x^T A x - 2 y_0 a^T x \\
& \text{s.t.} & \|x\|^2 = s^2 \\
& & y_0^2 = 1
\end{array}
$$

After homogenization: $q^* =$

$$
\begin{aligned}
&= \max_t \min_{\|x\|^2=s^2, y_0^2=1} x^T A x - 2y_0 a^T x + t(y_0^2 - 1) \\
&\geq \max_t \min_{\|x\|^2+y_0^2=s^2+1} x^T A x - 2y_0 a^T x + t(y_0^2 - 1) \\
&\quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&= \sup_\lambda \min_{x, y_0^2=1} x^T A x - 2y_0 a^T x + \lambda(\|x\|^2 - s^2) \\
&= \min_{x, y_0^2=1} \sup_\lambda x^T A x - 2y_0 a^T x + \lambda(\|x\|^2 - s^2) \\
&= \min_{\|x\|^2=s^2, y_0^2=1} x^T A x - 2y_0 a^T x \\
&= q^*.
\end{aligned}
$$

All of the above are equal, and

$$q^* = \max_t \min_{\|x\|^2 + y_0^2 = s^2+1} x^T A x - 2 y_0 a^T x + t(y_0^2 - 1)$$

$$= \max_t \min_{\|z\|^2 = s^2+1} z^T D(t) z - t = \max_t (s^2+1)\lambda_1(D(t)) - t$$

where $z = \begin{pmatrix} y_0 \\ x \end{pmatrix}$ and $D(t) = \begin{pmatrix} t & -a^T \\ -a & A \end{pmatrix}$.

- Rayleigh Quotient Used: For symmetric $G \in \mathcal{S}^n$ we have

$$\lambda_{\min}(G) = \min_{x \neq 0} \frac{x^T G x}{x^T x}.$$

Define

$$\boxed{k(t) := (s^2+1)\lambda_1(D(t)) - t}$$

Unconstrained dual problem for TRS is

$$\max_{t\in\mathbb{R}} k(t), \quad \text{(concave, coercive)}$$

rewrite into a linear semidefinite program:

$$\max_{D(t)\succeq\lambda I} (s^2+1)\lambda - t$$

Primal-Dual Pair of SDPs

$$
\begin{array}{rll}
q^* = & \max & (s^2+1)\lambda - t \\
& \text{s.t.} & \lambda I - tE_{00} \preceq D(0),
\end{array}
$$

$$
\begin{array}{rll}
q^* = & \min & \operatorname{trace} D(0)Y \\
& \text{s.t.} & \operatorname{trace} Y = s^2+1 \\
& & -Y_{00} = -1 \\
& & Y \succeq 0.
\end{array}
$$

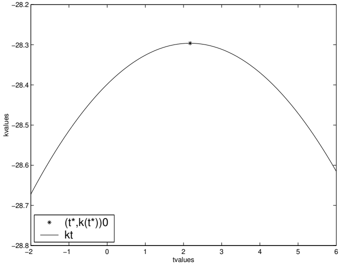
(Slater's CQ - strict complementarity - stable problems!?)

RW Algor. 1997, Modified

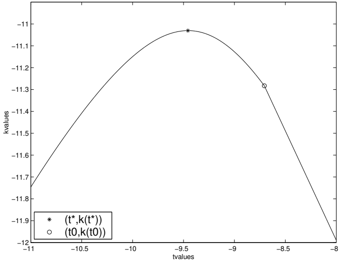
Rendl-Wolkowicz 1997 (large scale TRS) uses max-min eigenvalue problem $\max_t k(t) = (s^2+1)\lambda_1(D(t)) - t$
Define:

$$t_0 := \lambda_1(A) + \sum_{j \in \{k | (P^T a)_k \neq 0\}} \frac{(P^T a)_j^2}{\lambda_j(A) - \lambda_1(A)}.$$

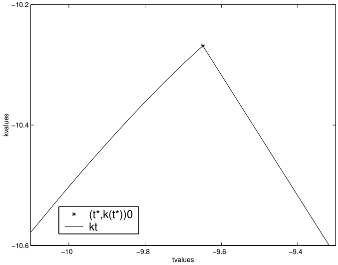
$k(t)$ in Easy Case



$k(t)$ in Hard Case (Case 1)



$k(t)$ in Hard Case (Case 2)



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

$y(t)$ normalized eigenvector for $\lambda_1(D(t))$

$$y(t) = \begin{pmatrix} y_0(t) \\ x(t) \end{pmatrix}$$

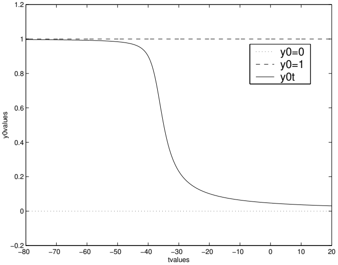
$$\frac{1}{y_0(t)}\|x(t)\| = s \text{ if and only if } k'(t) = 0$$

Properties of $k(t)$

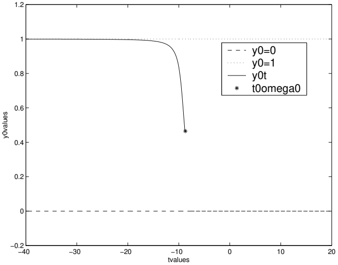
THEOREM: Let: $y(t)$ normalized eigenvector for $\lambda_1(D(t))$; $y_0(t)$ its first component. Then:

1. **In the easy case**: for $t \in \mathcal{R}$, $y_0(t) \neq 0$;
2. **In the hard case**:
   1. **for** $t < t_0$: $y_0(t) \neq 0$;
   2. **for** $t > t_0$: $y_0(t) = 0$;
   3. **for** $t = t_0$: $\exists$ basis eigenvectors for $\lambda_1(D(t_0))$, s.t. one, ω, has first component ($\omega_0 \neq 0$); and other eigenvectors have zero first component ($y_0(t) = 0$). □

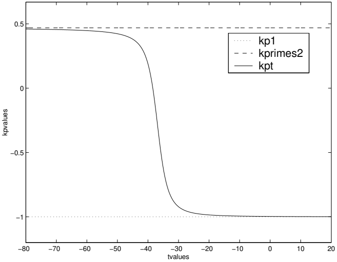
$y_0(t)$ in Easy Case



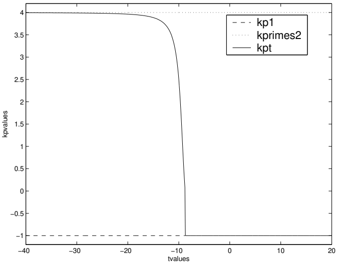
$y_0(t)$ in the Hard Case



$k'(t)$ in the Easy Case



$k'(t)$ in the hard case (*case 1*)

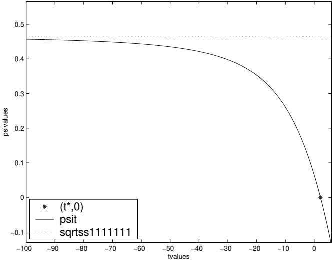


$\psi(t)$ in the Easy Case

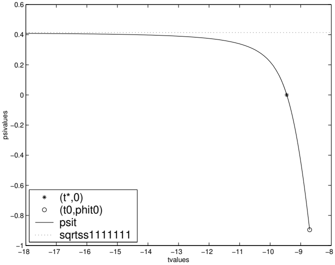
$\psi(t) = \sqrt{s^2+1} - \frac{1}{y_0(t)}$

Solving

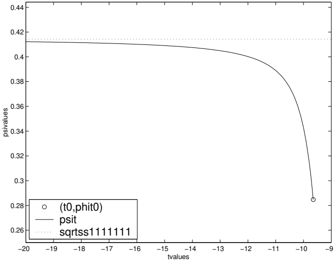
$\psi(t) = 0$; equivalently solving $k'(t) = 0$



$\psi(t)$ in the Hard Case (*case 1*)



$\psi(t)$ in the hard case (*case 2*)



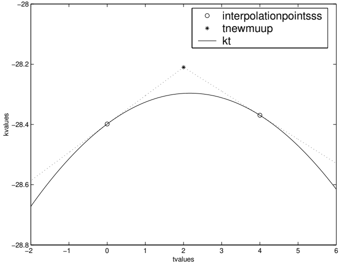
Newton Method on Moving Target

Newton Method on: $k(t) - M_t = 0$

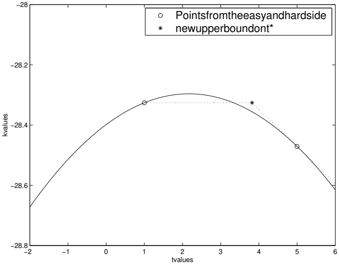
$$
\begin{aligned}
t_+ &= t_c - \frac{k(t_c) - M_t}{k'(t_c)} \\
&= \frac{(s^2+1)(t_c y_0^2(t_c) - \lambda(D(t_c)) - M_t}{(s^2+1) y_0(t_c)^2 - 1}
\end{aligned}
$$

Triangle Interpolation using $k(t)$

Reduce interval of uncertainty of t; improve upper bounds for q^*



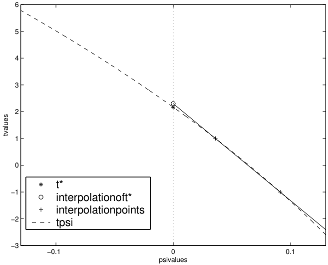
Vertical Cut



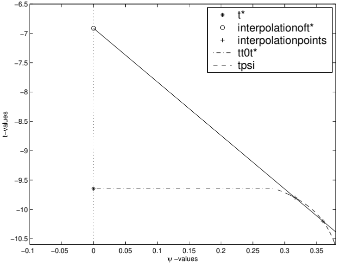
Inverse Interpolation

$$\begin{bmatrix} \psi_1^2 & \psi_1 & 1 \\ \psi_2^2 & \psi_2 & 1 \\ \psi_3^2 & \psi_3 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ t_{\text{new}} \end{bmatrix} = \begin{bmatrix} t_1 \\ t_2 \\ t_3 \end{bmatrix}$$

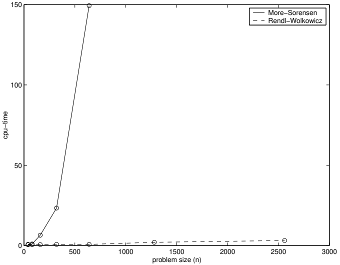
Inverse interpolation; easy cases



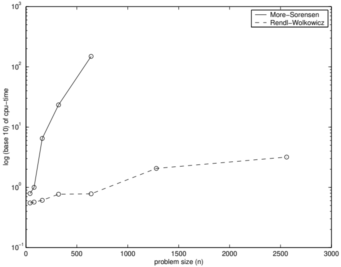
Inverse interpolation hard case (*case 2*)



cputime in the hard case (*case 2*)



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



Conclusion

Advantages:

- Hard Case is changed/equivalent to Linear Least Squares Problem
- Duality/geometry used for fast convergence/high accuracy solutions
- Large/Huge sparse problems solved quickly to high accuracy

Difficulties:

- Accurate computation of the smallest eigenvalue and corresponding eigenvector, e.g. in the case of large scale problems with multiple eigenvalues.