The Trust Region Subproblem and Regularization

Henry Wolkowicz

Department of Combinatorics and Optimization

University of Waterloo

(with Fortin and Grodzevich)

Sample Contributions

- G.E. Forsythe and G.H. Golub. On the stationary values of a second-degree polynomial on the unit sphere. *J. Soc. Indust. Appl. Math.*, 13:1050–1068, 1965.
- W. Gander, G.H. Golub, and U. von Matt. A constrained eigenvalue problem. *Linear Algebra Appl.*, 114/115:815–839, 1989.
- G. Golub and U. von Matt. Quadratically constrained least squares and quadratic problems. *Numer. Math.*, 59:561–580, 1991.
- G. H. Golub and U. von Matt. Tikhonov regularization for large scale problems. In *Scientific computing (Hong Kong, 1997)*, pages 3–26. Springer, Singapore, 1997.

The Book

“How did the Golub and Van Loan book come about? Roger Horn was the founder of the Department of Mathematical Sciences at Johns Hopkins. In cooperation with Johns Hopkins Press, he held a series of short courses each documented in a monograph. He had invited Gene to teach one of these courses. Charlie Van Loan was there. They decided to try to write a monograph.” (sample attendees: George Styan, Pete Stewart, Richard Bartels,...H.W.)

Unconstrained Minimization

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

Quadratic Model at current estimate x_c:

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

The optimal d exists and can be found efficiently.

The Trust Region Subproblem

(TRS)

$$q^* = \min_x \quad q(x) := x^T A x - 2a^T x$$
$$\text{s.t.} \quad \|x\| \le s, x \in \mathbb{R}^n$$

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

LLS $\min_x \|Gx - d\|_2^2, \quad G$ 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 $\|x(\hat{s})\| \leq \hat{s}$

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

x^* optimal for TRS

if and only if

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

dual feas.

primal feas.

compl. slack.

(phrased in the modern primal-dual paradigm)

Surprising: characterization of opt.; 2nd order psd

Duality and MS Algorithm

- Hidden constraint
- lower bound yields dual (SDP)
- Convexity/SDP Arise Naturally
- i.e. dual algor: maintain dual feasibility while trying to attain primal feasibility

Background

Regularization: linear least-squares problems

$$\textbf{LLS} \qquad \min_x \|Gx - d\|_2,$$

$G, n \times n$ is: singular or ill-conditioned forward operator; d is: observed data; with noise η:

$$Gx = Gx_{\text{true}} + \eta = d = d_{\text{true}} + \eta.$$

remarkable fact: for many applications, a small amount of noise η can result in a solution $x = G^{\dagger}d$ that has no relation to x_{true}

Regularization (TRS, Tikhonov)

AIM: *find generalized solutions* stable under small changes in d

TRS Approach:

$$\textbf{TRS} \qquad \min \|Gx - d\|_2^2 \quad \text{s.t.} \|x\|_2^2 \leq \varepsilon^2$$

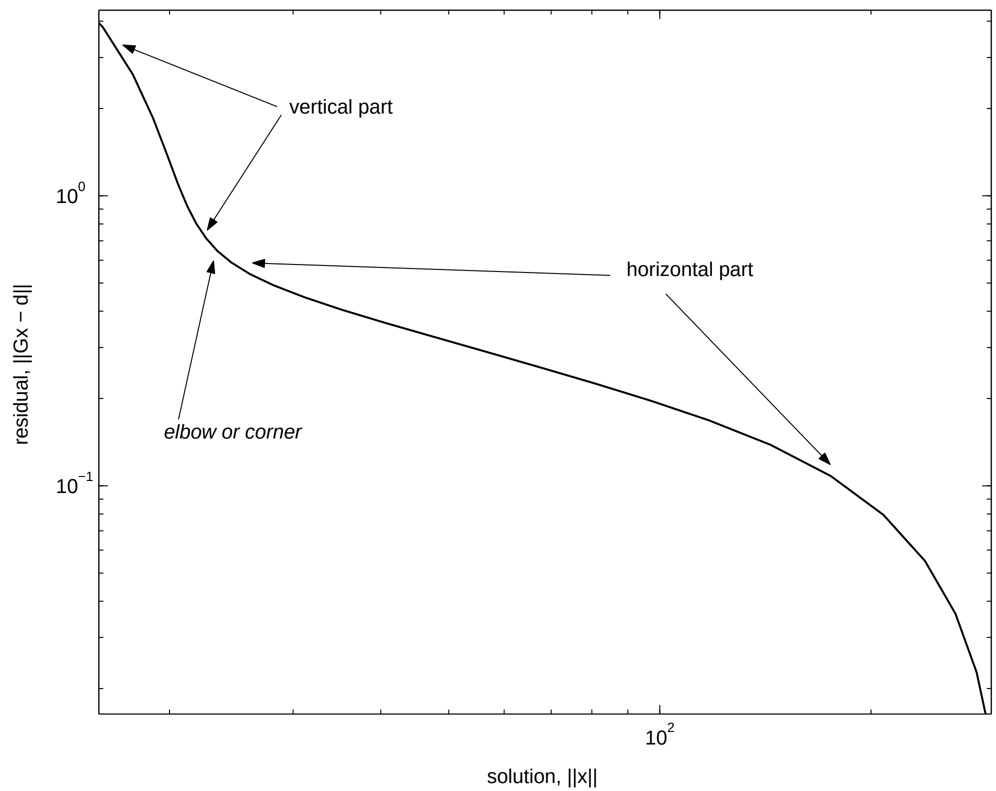
Tikhonov Regularization:

$$\textbf{TikhReg} \qquad (G^T G + \alpha^2 I)x_\alpha = G^T d$$

Result: larger residual error $\|Gx - d\|_2$ but smaller propagated error in $\|x\|_2$.

L-Curve (Point of Max Curvature/elbow)

$\mathcal{L}(G,d) = \{(\log(\varepsilon), \log \|Gx(\varepsilon) - d\|_2) : \quad \varepsilon > 0,\ x(\varepsilon)$ optimal for $TRS\}$.



reformulate as TRS

$$\mu_\varepsilon := \mu(A, a, \varepsilon) := \begin{array}{ll} \min & q(x) := x^T A x - 2a^T x \\ \text{s.t.} & \|x\|_2^2 \leq \varepsilon^2, \end{array}$$

$A = G^T G$ (nonsingular, ill-cond.)
$a = G^T d \in \mathbb{R}^n$, $\varepsilon > 0$, $x \in \mathbb{R}^n$

Curvature of the L-curve

$$\eta = \|x_\varepsilon\|_2^2, \quad \hat{\eta} = \log \eta; \qquad \rho = \|Gx_\varepsilon - d\|_2^2 = \mu_\varepsilon + d^T$$

curvature of L-curve

$$\kappa_\varepsilon = 2\frac{\hat{\rho}'\hat{\eta}'' - \hat{\rho}''\hat{\eta}'}{((\hat{\rho}')^2 + (\hat{\eta}')^2)^{3/2}}$$

$$= \varepsilon^2 \mu_\varepsilon \left(2\varepsilon^2 \lambda^{*2} - 2\mu_\varepsilon \lambda^* - \varepsilon \mu_\varepsilon \left(\frac{\partial \lambda^*}{\partial \varepsilon}\right)\right)\left(\varepsilon^4 \lambda^{*2} + \mu\right.$$

$\frac{\partial \lambda^*}{\partial \varepsilon} = \varepsilon / (a^T (A - \lambda^* I)^{-3} a)$ expensive

Curvature Estimation; Gauss Quadrature

Denominator of $\frac{\partial \lambda^*}{\partial \varepsilon} = \varepsilon/(a^T(A - \lambda^* I)^{-3} a)$, is expensive

Use (ref. Golub and Von Matt): find upper and lower bounds

$$l_p(\alpha) \leq \nu_p(\alpha) = d^T G (G^T G + \alpha I)^p G^T d \leq u_p(\alpha)$$

where $\alpha > 0$; p negative integer

($\alpha = -\lambda^*, p = -3, G^T G = A, G^T d = a$)

- bounds from Lanczos Bidiagonalization on G (with restarts);
- accuracy can be increased as needed

Four Parameters

t – control parameter in $k(t), D(t)$
ε – $= \|x(\varepsilon)\|_2$, trust-region radius
α – Tikhonov regularization parameter
λ – optimal TRS Lagr. mult.

Relationships (isotonic):

$$
\begin{array}{rclcl}
-\infty & < & \lambda = \lambda_1(D(t)) = -\alpha^2 & \leq & 0 \\
0 & < & t = \lambda + d^T G(G^T G - \lambda I)^{-1} G^T d & \leq & \|d\|_2^2 \\
0 & < & \varepsilon = \|(G^T G - \lambda I)^{-1} G^T d\|_2 & \leq & \|G^{-1}d
\end{array}
$$

Upper bound corresponds to the LLSS

Geometry of Elbow

Define

$$l_r(\varepsilon) := \log(\|Gx(\varepsilon) - d\|_2), \quad l_x(\varepsilon) := \log(\|x(\varepsilon)\|_2)$$

$$\begin{aligned} \frac{\partial(l_r(\varepsilon))}{\partial(l_x(\varepsilon))} &= \frac{1}{2}\frac{\partial(\log(\mu_\varepsilon + d^T d))/\partial(\varepsilon)}{\partial(\log(\varepsilon))/\partial(\varepsilon)} \\ &= \frac{1}{2}\frac{\mu_\varepsilon' \varepsilon}{\mu_\varepsilon + d^T d} \\ &= \frac{\varepsilon^2 \lambda_\varepsilon}{\mu_\varepsilon + d^T d} \end{aligned}$$

large negative number as we approach *elbow* from left

negative close to zero at the plateau right of *elbow*

Initial L-curve point

- We start to the left of the elbow;
- each iterate increases the value of t to locate the elbow
- exploit relationships between different parameters

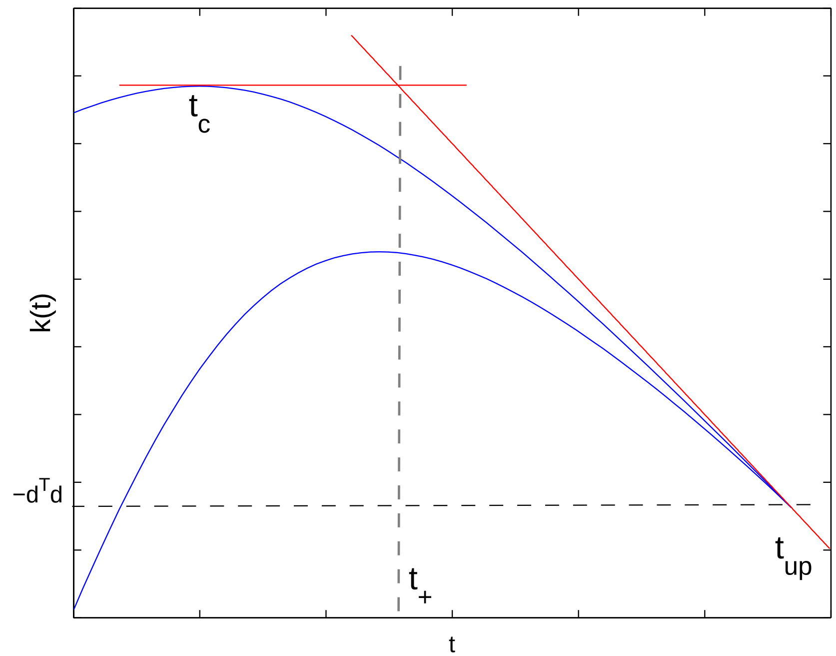
e.g. use initial point $\lambda = -\sigma_n(G)^2$ (lower bound on optimal λ^*) or use small enough value $t = \frac{d^T d}{2}$

Geometry of L-Curve

small changes in t result in large changes in ε on the horizontal part to the right (plateau) of the elbow

Conversely, large changes in t result in small changes in ε when on the vertical part to the left of the elbow.

Triangle Interpolation for t_+

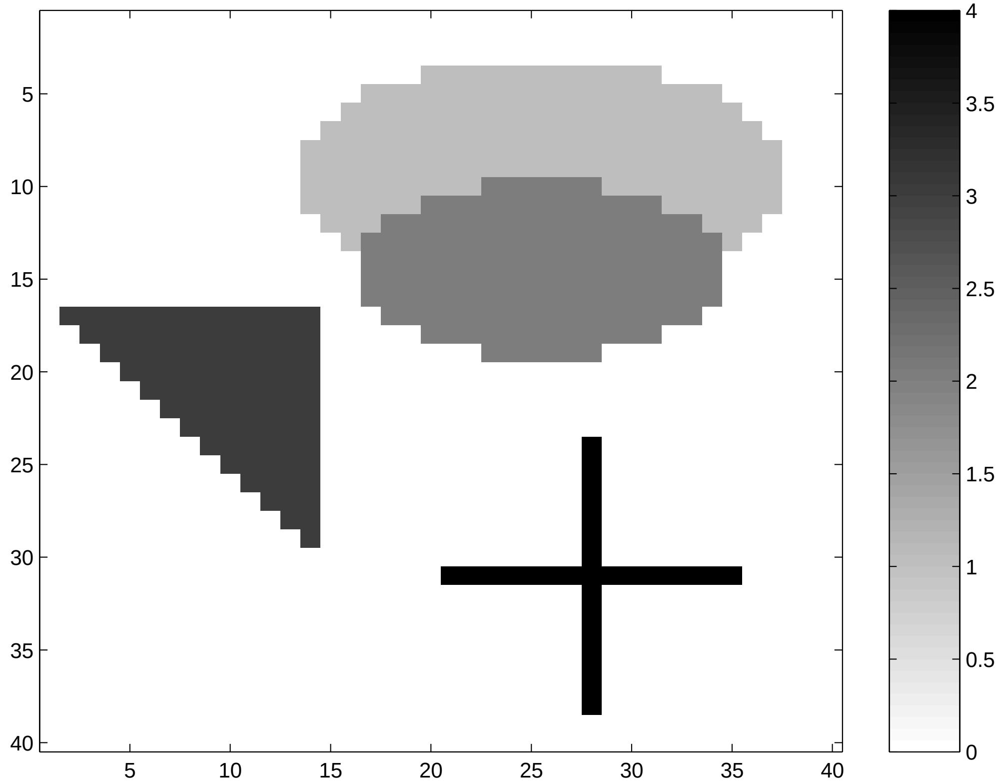


(str. concave)

$k(t) = (\varepsilon_c^2 + 1)\lambda_{\min}(D(t)) - t; \qquad \mu_\varepsilon = \max_t k(t)$

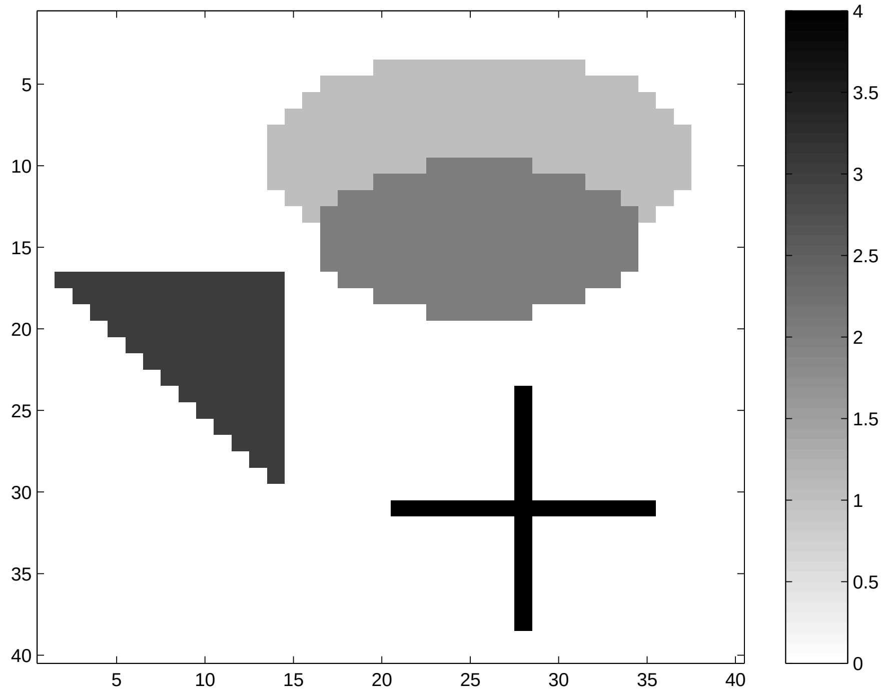
$k'(t_c) = 0$, $k'(t_{up}) = -1$, independent of ε

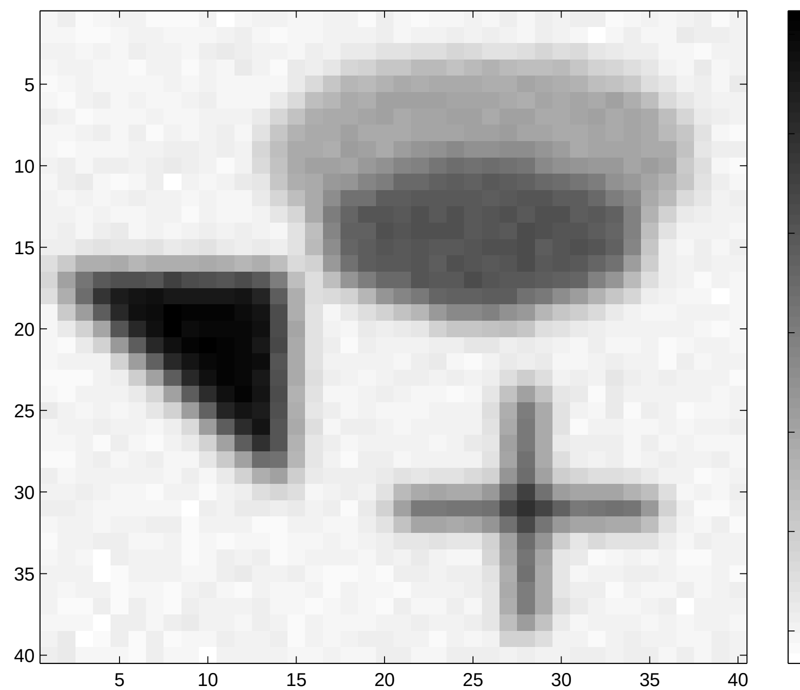
Original Picture



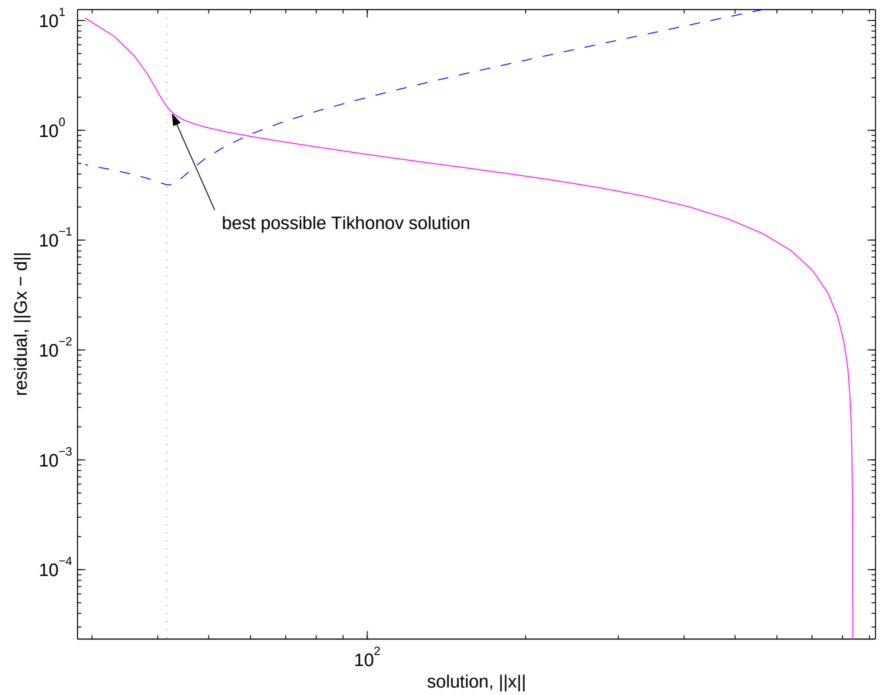
(ref. Hansen MATLAB package deblurring example - use blur command)

Original; Blurred/Noise





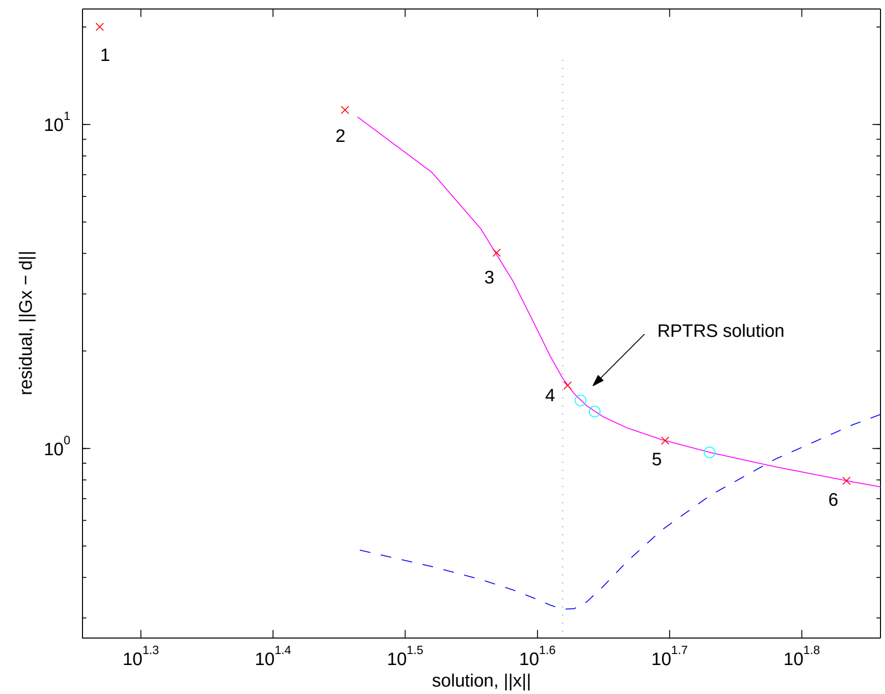
Corresponding L-curve



dashed line shows relative accuracy $\frac{\|x_{\text{true}} - x\|_2}{\|x_{\text{true}}\|_2}$

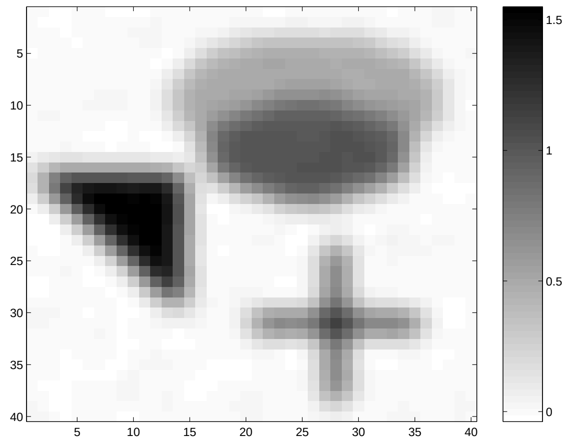
minimum - best possible Tikhonov regularization point

L-curve with RPTRS points

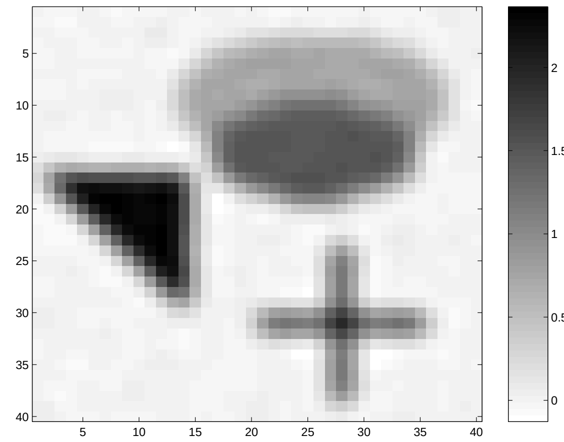


x (cross) visited during the main loop
circles final refinement steps
Final point - close to best Tikhonov solution!

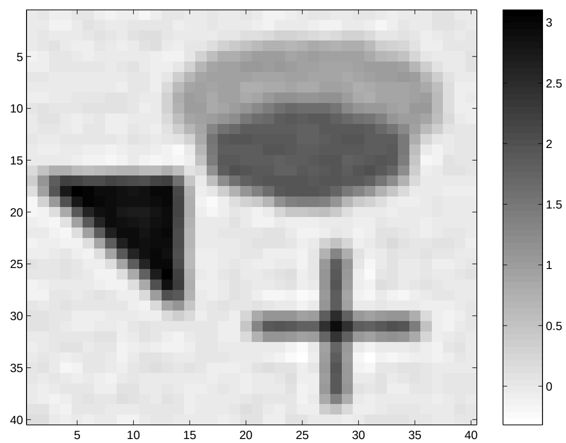
Points 1-4



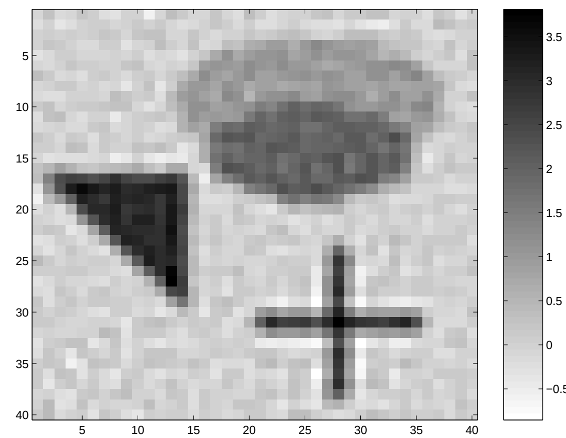
t = 652.166, rel.acc. = 65.39



t = 994.155, rel.acc. = 49.63

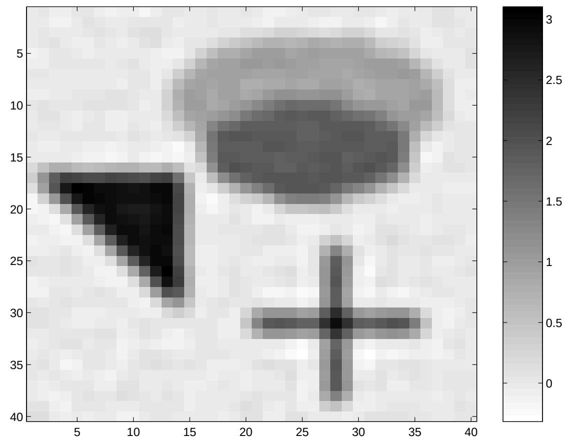


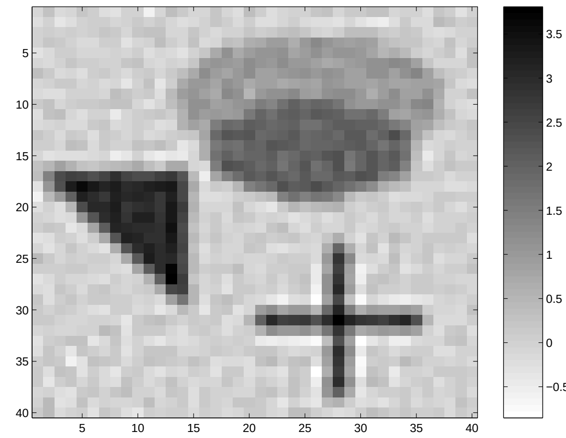
t = 1271.46, rel.acc. = 38.07



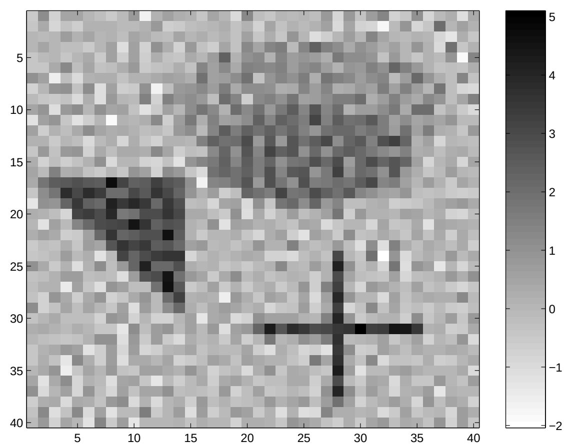
t = 1378.38, rel.acc. = 31.82

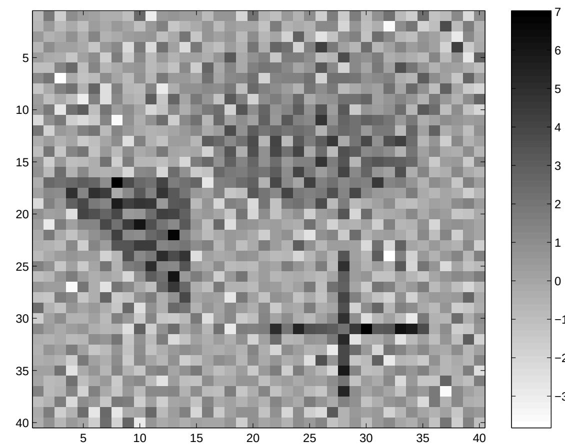
Points 3-6





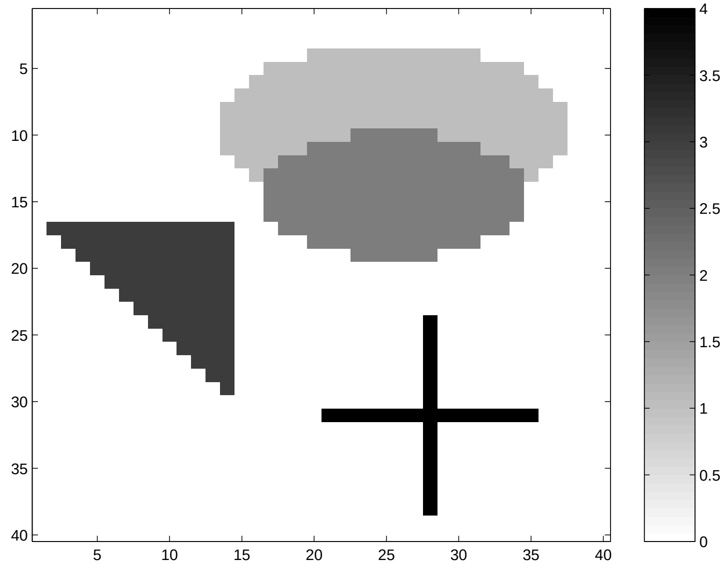
t = 1271.46, rel.acc. = 38.07   t = 1378.38, rel.acc. = 31.82

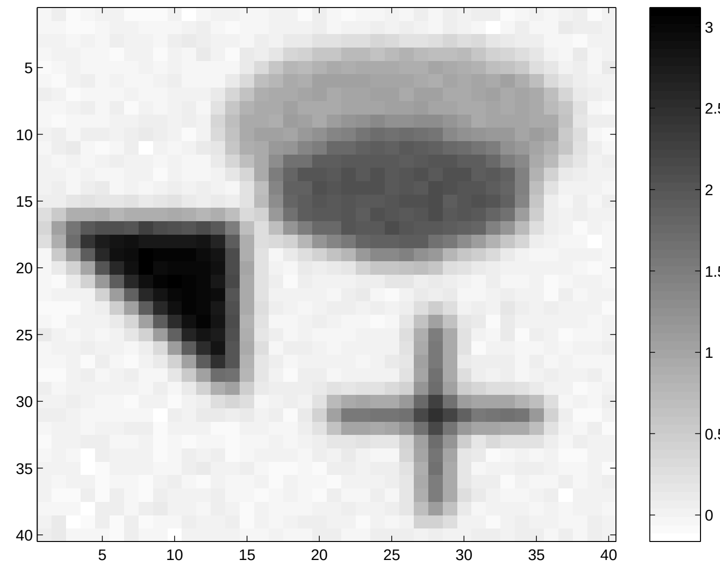


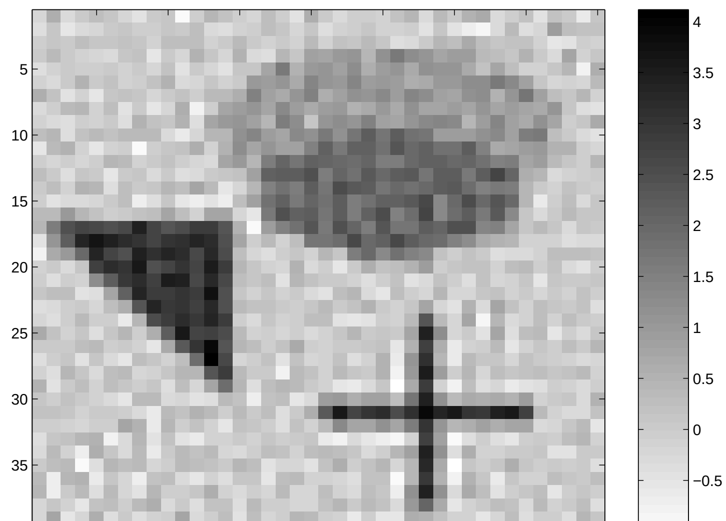


t = 1392.12, rel.acc. = 57.14   t = 1393.45, rel.acc. = 116.29

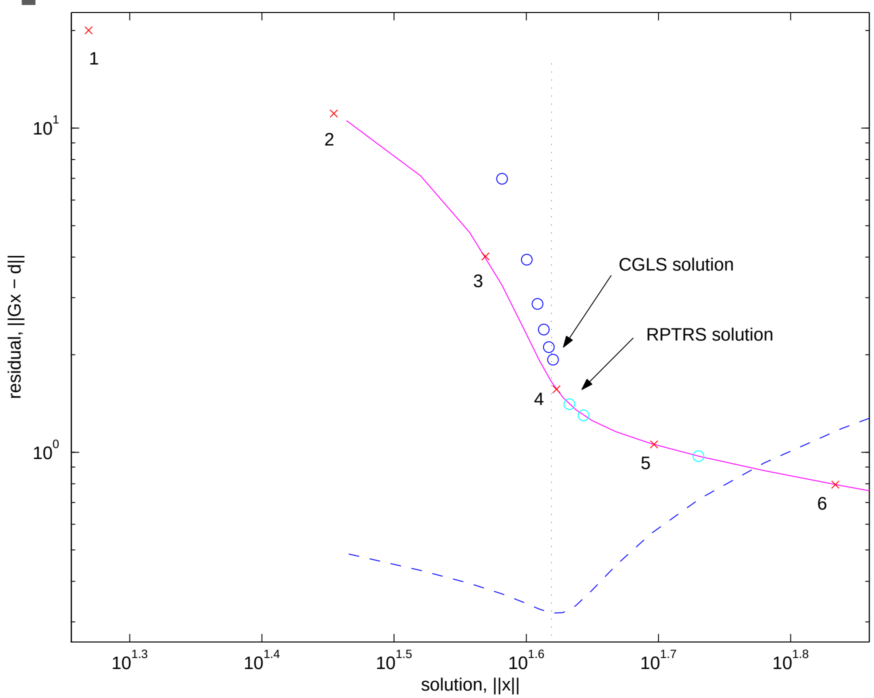
Original; Blurred; RPTRS Solution

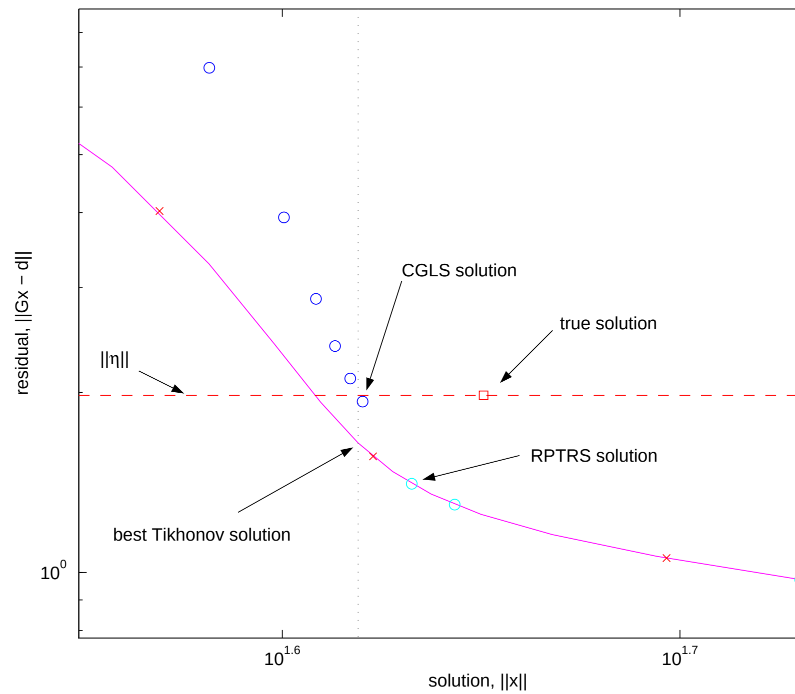






L-curve with CGLS and RPTRS





L-curve with CGLS best Tikhonov

CGLS, RPTRS with

Conclusion

- Used ideas from RW algorithm for TRS
- new algorithm efficiently finds point of maximum curvature on the L-curve for regularization of ill-conditioned problems $Gx = d$
- takes advantage of: each iteration of RW algorithm corresponds to a point on the L-curve; cost is approx. ONE TRS solve
- Advantage over CGLS approach when the norm of the error is not known

References

[1] G.E. FORSYTHE and G.H. GOLUB. On the stationary values of a second-degree polynomial on the unit sphere. *J. Soc. Indust. Appl. Math.*, 13:1050–1068, 1965.