

Choosing the Tikhonov regularization parameter in large scale problems

G. Rodriguez

in collab. with Lothar Reichel and Sebastiano Seatzu

Dipartimento di Matematica e Informatica, Università di Cagliari
viale Merello 92, 09123 Cagliari, Italy

Structured Numerical Linear Algebra Problems:
Algorithms and Applications
Cortona - September 15–19, 2008

Error estimates for choosing a regularization parameter

In two recent papers, Brezinski, Seatzu and R developed various classes of error estimates for linear systems and applied them to the selection of the parameter in various regularization techniques.

We will describe a method for selecting the Tikhonov regularization parameter in large scale problems and compute the regularized solution.

This procedure is based on the minimization of an error estimate and requires at each step two matrix-vector products (Az and A^Tz).

Error estimates for rectangular linear systems

Let $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) = r \leq \min(m, n)$ and

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|A\mathbf{x} - \mathbf{b}\|_2, \quad \mathbf{b} = \hat{\mathbf{b}} + \sigma \|\hat{\mathbf{b}}\| \boldsymbol{\varepsilon}.$$

The **normal solution** $\mathbf{x}^\dagger = A^\dagger \mathbf{b}$ solves the minimization problem

$$\min_{\mathbf{x} \in \mathcal{S}} \|\mathbf{x}\|_2, \quad \mathcal{S} = \{\mathbf{x} \in \mathbb{R}^n : A^T(\mathbf{b} - A\mathbf{x}) = 0\}.$$

When \mathbf{x} is an approximate solution, $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ and $\mathbf{e} = \mathbf{x}^\dagger - \mathbf{x}$, [BRS] showed that

$$\|\mathbf{e}\|^2 \simeq \boxed{\eta_\nu^2 = d_0^{\nu-1} d_1^{5-2\nu} d_2^{\nu-3}} \quad \nu \in \mathbb{R}$$

where

$$d_0 = \|\mathbf{r}\|^2, \quad d_1 = \|A^T \mathbf{r}\|^2, \quad d_2 = \|AA^T \mathbf{r}\|^2.$$

Tikhonov regularization

$$J(\mathbf{x}, \mu) = \|A\mathbf{x} - \mathbf{b}\|^2 + \mu\|\mathbf{x}\|^2 \quad \rightarrow \quad \mathbf{x}_\mu = \arg \min_{\mathbf{x} \in \mathbb{R}^n} J(\mathbf{x}, \mu)$$

The regularized vector \mathbf{x}_μ is the solution of the **normal equation**

$$(A^T A + \mu I_n) \mathbf{x}_\mu = A^T \mathbf{b}.$$

This equation implies

$$A^T \mathbf{r}_\mu = A^T (\mathbf{b} - A\mathbf{x}_\mu) = \mu \cdot \mathbf{x}_\mu,$$

so that

$$\boxed{\eta_\nu(\mu) = \|\mathbf{r}_\mu\|^{\nu-1} \cdot \|\mathbf{x}_\mu\|^{5-2\nu} \cdot \|A\mathbf{x}_\mu\|^{\nu-3} \cdot \mu^{2-\nu}} \quad \nu \in \mathbb{R}.$$

Representation of residual

Introducing the spectral decomposition $AA^T = W\Lambda W^T$ in

$$\|\mathbf{r}_\mu\|^2 = \mu^2 \mathbf{b}^T (AA^T + \mu I)^{-2} \mathbf{b}$$

we obtain

$$\|\mathbf{r}_\mu\|^2 = \mu^2 \mathbf{w}^T (\Lambda + \mu I)^{-2} \mathbf{w} = \sum_{j=1}^m \phi_\mu(\lambda_j) w_j^2,$$

where $\mathbf{w} = W^T \mathbf{b}$ and $\phi_\mu(t) := \left(\frac{\mu}{t+\mu}\right)^2$.

Employing Stieltjes integrals and Gauss quadrature

The main idea [Golub et al.] consists of three steps.

- 1 Set

$$\|\mathbf{r}_\mu\|^2 = \sum_{j=1}^m \phi_\mu(\lambda_j) w_j^2 = \int_{-\infty}^{\infty} \phi_\mu(t) dw(t),$$

where $dw(t)$ is a nonnegative measure, whose distribution function is nondecreasing and piecewise constant, with jumps at the eigenvalues λ_j of AA^T .

- 2 Construct upper and lower bounds for the integral by means of **Gauss quadrature** rules.
- 3 Perform the computation via the **Lanczos bidiagonalization** algorithm.

Stieltjes integral braketing

From the remainder in a GQ formula with m prescribed nodes

$$R[f] = \frac{f^{(2n+m)}(\xi)}{(2n+m)!} \int_a^b \prod_{k=1}^m (t - z_k) \left[\prod_{j=1}^n (t - t_j) \right]^2 dw(t),$$

since $\phi_\mu^{(2k)}(t) > 0$ for $t > 0$, we get

$$\mathcal{G}_{\ell-1}(\phi_\mu) < \mathcal{G}_\ell(\phi_\mu) < \|\mathbf{r}_\mu\|^2 < \mathcal{R}_k(\phi_\mu) < \mathcal{R}_{k-1}(\phi_\mu)$$

where

\mathcal{G}_ℓ ℓ -point Gauss quadrature rule

\mathcal{R}_k k -point Gauss-Radau quadrature rule
with a prescribed node at the origin

Lanczos bidiagonalization

Fixed $\mathbf{u}_1 = \mathbf{b}/\|\mathbf{b}\|$, it consists of computing iteratively

$$AV_\ell = U_{\ell+1} \bar{C}_\ell, \quad \ell = 1, 2, \dots$$

where the columns of V_ℓ and $U_{\ell+1}$ are orthogonal and

$$\bar{C}_\ell = \begin{bmatrix} \rho_1 & & & 0 \\ \sigma_2 & \rho_2 & & \\ & \ddots & \ddots & \\ 0 & & \sigma_\ell & \rho_\ell \\ & & & \sigma_{\ell+1} \end{bmatrix} = \begin{bmatrix} C_\ell \\ \sigma_{\ell+1} \mathbf{e}_\ell^T \end{bmatrix}.$$

Gauss quadrature & Lanczos bidiagonalization

We obtain

$$AA^T U_\ell = U_\ell C_\ell C_\ell^T + \sigma_{\ell+1} \rho_\ell \mathbf{u}_{\ell+1} \mathbf{e}_\ell^T$$

and

$$\mathbf{u}_j = \rho_{j-1}(AA^T)\mathbf{b}, \quad 1 \leq j \leq \ell + 1.$$

for a set of polynomials $p_{j-1} \in \Pi_{j-1}$ **orthogonal** w.r. to

$$\langle f, g \rangle = \int_{-\infty}^{\infty} f(t)g(t) dw(t)$$

whose **recursion coefficients** are contained in $T_\ell = C_\ell C_\ell^T$.

Gauss quadrature & Lanczos bidiagonalization

This leads to

$$\begin{aligned}\mathcal{G}_\ell(\phi_\mu) &= \mu^2 \|\mathbf{b}\|^2 \cdot \mathbf{e}_1^T (C_\ell C_\ell^T + \mu I) \mathbf{e}_1 \\ \mathcal{R}_{\ell+1}(\phi_\mu) &= \mu^2 \|\mathbf{b}\|^2 \cdot \mathbf{e}_1^T (\bar{C}_\ell \bar{C}_\ell^T + \mu I) \mathbf{e}_1.\end{aligned}$$

Writing

$$\mathcal{G}_\ell(\phi_\mu) = \|\mathbf{b}\|^2 \cdot \mathbf{y}_\ell^T(\mu) \mathbf{y}_\ell(\mu),$$

the vector $\mathbf{y}_\ell(\mu) := \mu(C_\ell C_\ell^T + \mu I)^{-1} \mathbf{e}_1$ is the solution of the least-squares problem

$$\min_{\mathbf{y} \in \mathbb{R}^\ell} \left\| \begin{bmatrix} C_\ell^T \\ \mu^{1/2} I_\ell \end{bmatrix} \mathbf{y} - \mu^{1/2} \mathbf{e}_{\ell+1} \right\|;$$

see [Elden (1977)] for an algorithm.

The rest of the computation...

The same idea can be applied to the other factors of $\eta_\nu(\mu)$:

$$\|\mathbf{x}_\mu\|^2 = \hat{\mathbf{w}}^T (\hat{\Lambda} + \mu I)^{-2} \hat{\mathbf{w}} = \int_{-\infty}^{\infty} \frac{1}{\mu^2} \phi_\mu(t) d\hat{w}(t),$$

with $A^T A = \hat{W} \hat{\Lambda} \hat{W}^T$ and $\hat{\mathbf{w}} = \hat{W}^T A^T \mathbf{b}$;

$$\mu^2 \|A\mathbf{x}_\mu\|^2 = \mu^2 \check{\mathbf{w}}^T (\Lambda + \mu I)^{-2} \check{\mathbf{w}} = \int_{-\infty}^{\infty} \phi_\mu(t) d\check{w}(t),$$

with $\check{\mathbf{w}} = W^T A A^T \mathbf{b}$.

Both integrals can be bracketed by GQ rules

$$\hat{\mathcal{G}}_{\ell-1}(\phi_\mu) < \hat{\mathcal{G}}_\ell(\phi_\mu) < \|\mathbf{x}_\mu\|^2 < \hat{\mathcal{R}}_k(\phi_\mu) < \hat{\mathcal{R}}_{k-1}(\phi_\mu),$$

$$\check{\mathcal{G}}_{\ell-1}(\phi_\mu) < \check{\mathcal{G}}_\ell(\phi_\mu) < \|A\mathbf{x}_\mu\|^2 < \check{\mathcal{R}}_k(\phi_\mu) < \check{\mathcal{R}}_{k-1}(\phi_\mu).$$

The rest of the computation . . .

We obtain the following bounds for $\|\mathbf{x}_\mu\|^2$:

$$\hat{\mathcal{G}}_\ell(\phi_\mu) = \|A^T \mathbf{b}\|^2 \mathbf{e}_1^T (\hat{C}_\ell \hat{C}_\ell^T + \mu I)^{-2} \mathbf{e}_1,$$

$$\hat{\mathcal{R}}_\ell(\phi_\mu) = \|A^T \mathbf{b}\|^2 \mathbf{e}_1^T (\bar{C}_{\ell-1} (\bar{C}_{\ell-1})^T + \mu I)^{-2} \mathbf{e}_1,$$

with

$$\|A^T \mathbf{b}\|^2 = \|\mathbf{b}\|^2 \rho_1^2,$$

$$\bar{C}_\ell = \bar{Q}_\ell \hat{C}_\ell^T,$$

$$\hat{C}_\ell = \begin{bmatrix} \bar{C}_{\ell-1} & \alpha_\ell \mathbf{e}_\ell \end{bmatrix}.$$

The rest of the computation . . .

We obtain the following bounds for $\|A\mathbf{x}_\mu\|^2$:

$$\begin{aligned}\check{\mathcal{G}}_{\ell-1}(f) &= \|AA^T \mathbf{b}\|^2 \mathbf{e}_1^T (\check{R}_{\ell-1}^T \check{R}_{\ell-1} + \mu I)^{-2} \mathbf{e}_1, \\ \check{\mathcal{R}}_{\ell-1}(f) &= \|AA^T \mathbf{b}\|^2 \mathbf{e}_1^T (\check{R}_{\ell-1,0}^T \check{R}_{\ell-1,0} + \mu I)^{-2} \mathbf{e}_1,\end{aligned}$$

with

$$\|AA^T \mathbf{b}\|^2 = \|\mathbf{b}\|^2 \rho_1^2 (\rho_1^2 + \sigma_2^2),$$

$$C_\ell = Q'_\ell R'_\ell,$$

$$(R'_\ell)^T = Q''_\ell R''_\ell,$$

$$\check{R}_{\ell-1} = R''_\ell (1 : \ell - 1, 1 : \ell - 1),$$

$$\check{R}_{\ell-1} = \begin{bmatrix} \check{R}_{\ell-1,0} \\ \beta_{\ell-1} \mathbf{e}_{\ell-1}^T \end{bmatrix}.$$

Bounds for error estimates

This procedure allows us to obtain **upper** and **lower bounds** for $\eta_\nu(\mu)$. For example:

$$\eta_2(\mu) = \frac{\|\mathbf{r}_\mu\| \cdot \|\mathbf{x}_\mu\|}{\|A\mathbf{x}_\mu\|},$$

$$\left(\frac{\mathcal{G}_\ell(\phi_\mu) \hat{\mathcal{G}}_\ell(\phi_\mu)}{\check{\mathcal{R}}_{\ell-1}(\phi_\mu)} \right)^{1/2} < \eta_2(\mu) < \left(\frac{\mathcal{R}_{\ell+1}(\phi_\mu) \hat{\mathcal{R}}_\ell(\phi_\mu)}{\check{\mathcal{G}}_{\ell-1}(\phi_\mu)} \right)^{1/2}.$$

The computation is performed by the application of a judiciously chosen sequence of simple orthogonal transformations.

It requires only $O(\ell)$ additional *flops*, besides the load of the Lanczos bidiagonalization.

The numerical method: first phase

We start with a grid of q log-equispaced values of μ ,
 $\mathcal{N} := \{\mu_1, \dots, \mu_q\}$ and $\mathcal{C} := \emptyset$.

- (i) For μ_j in \mathcal{N} , compute $\eta_\nu^l(\mu_j)$, $\eta_\nu^u(\mu_j)$ and their average $\bar{\eta}_\nu(\mu_j)$.
- (ii) Determine the smallest k , such that

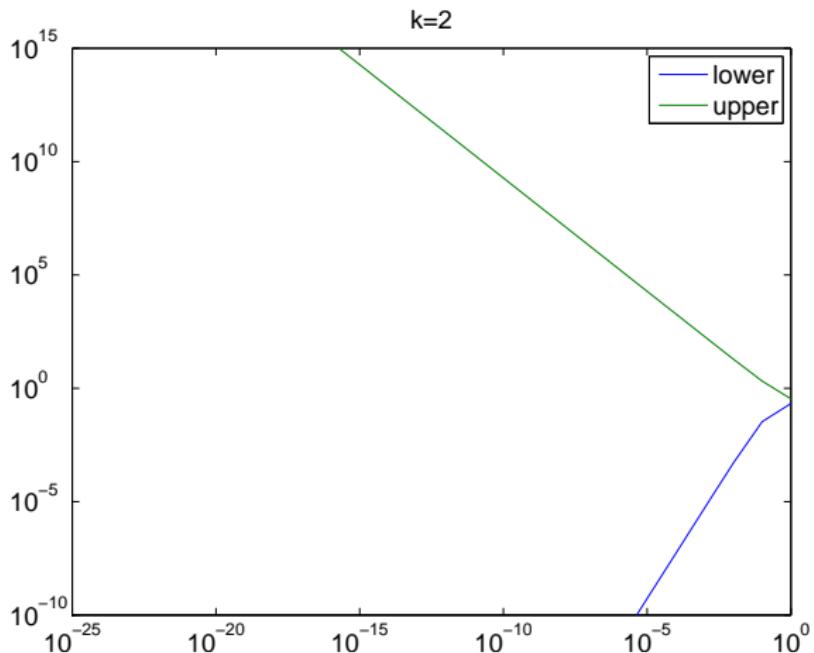
$$\left| \eta_\nu^u(\mu_j) - \eta_\nu^l(\mu_j) \right| < \beta \cdot \bar{\eta}_\nu(\mu_j), \quad j = k, k+1, \dots, \#\mathcal{N}.$$

- (iii) Move the parameters $\{\mu_k, \dots, \mu_{\#\mathcal{N}}\}$ from the set \mathcal{N} to the set \mathcal{C} .
- (iv) Repeat the computations of steps (i)-(iii) until

$$\bar{\eta}_\nu(\mu_k) > \bar{\eta}_\nu(\mu_{k+1})$$

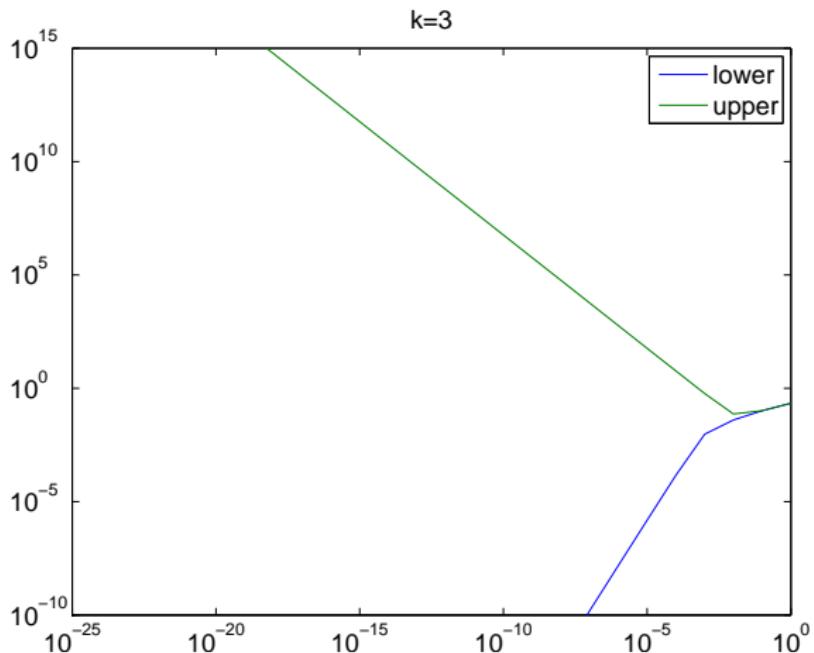
for some pair $\{\mu_k, \mu_{k+1}\} \subset \mathcal{C}$.

The algorithm at work: first phase



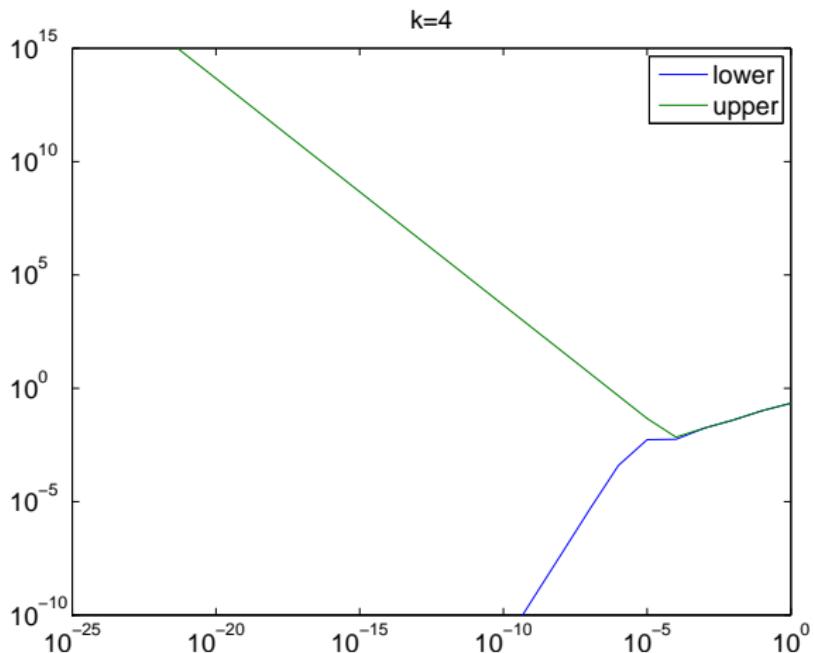
Baart test problem, $m = n = 200, \sigma = 10^{-4}$

The algorithm at work: first phase



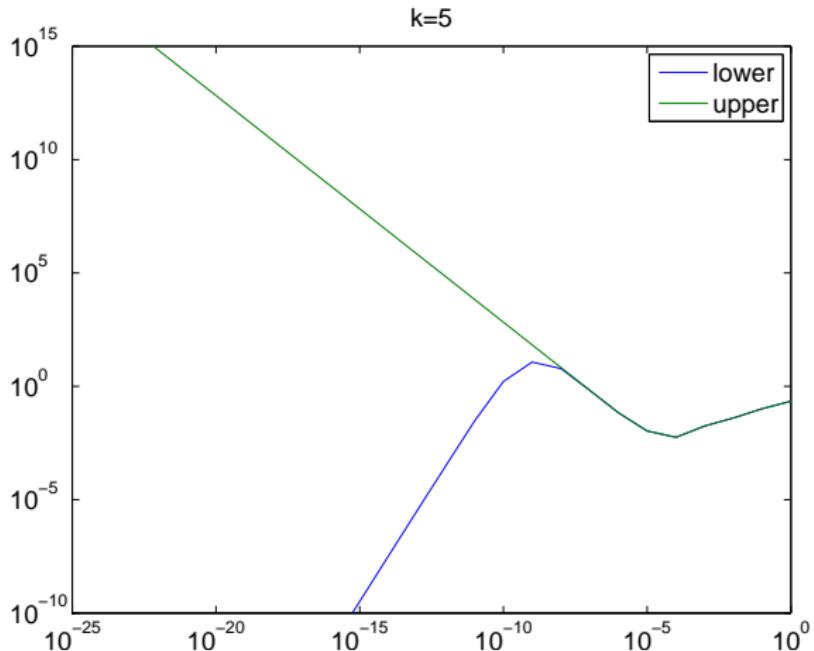
Baart test problem, $m = n = 200$, $\sigma = 10^{-4}$

The algorithm at work: first phase



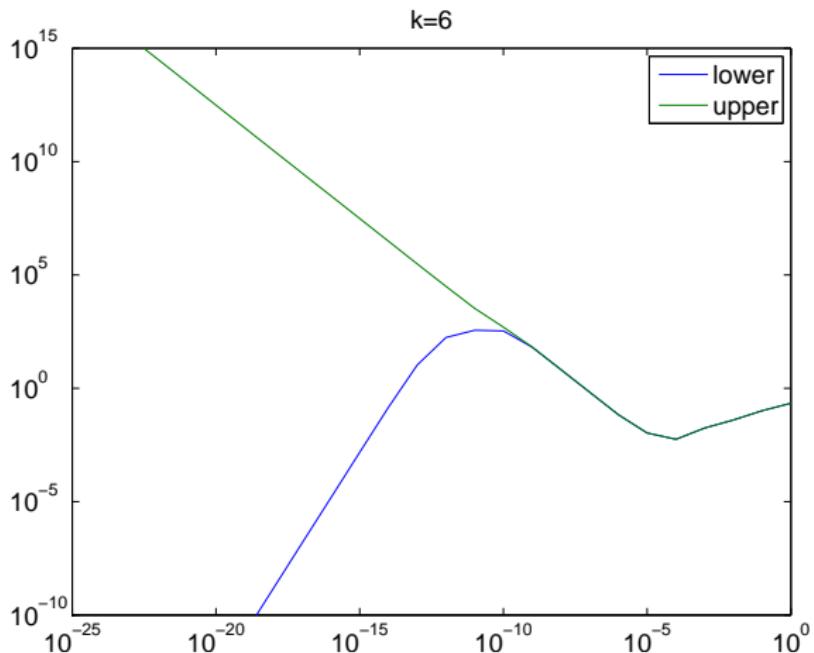
Baart test problem, $m = n = 200$, $\sigma = 10^{-4}$

The algorithm at work: first phase



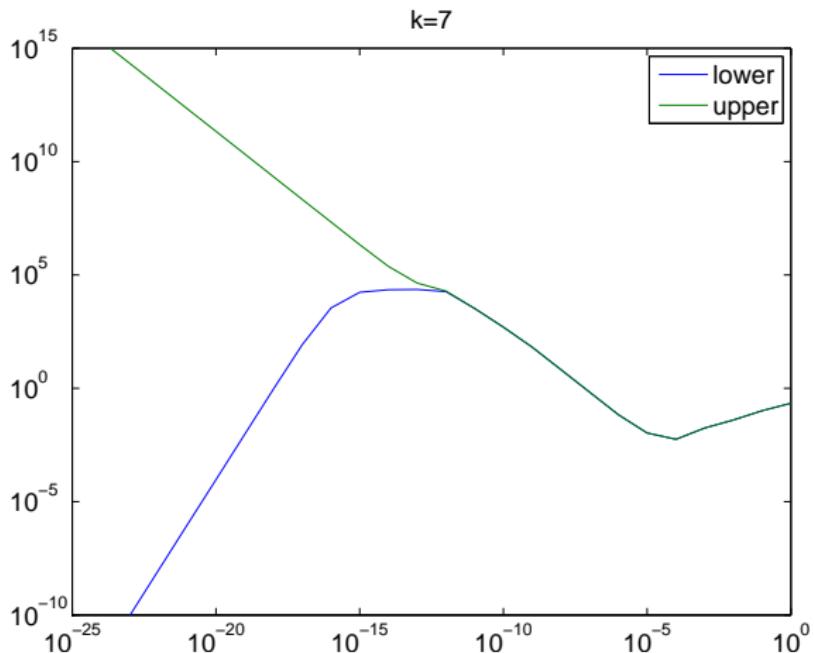
Baart test problem, $m = n = 200$, $\sigma = 10^{-4}$

The algorithm at work: first phase



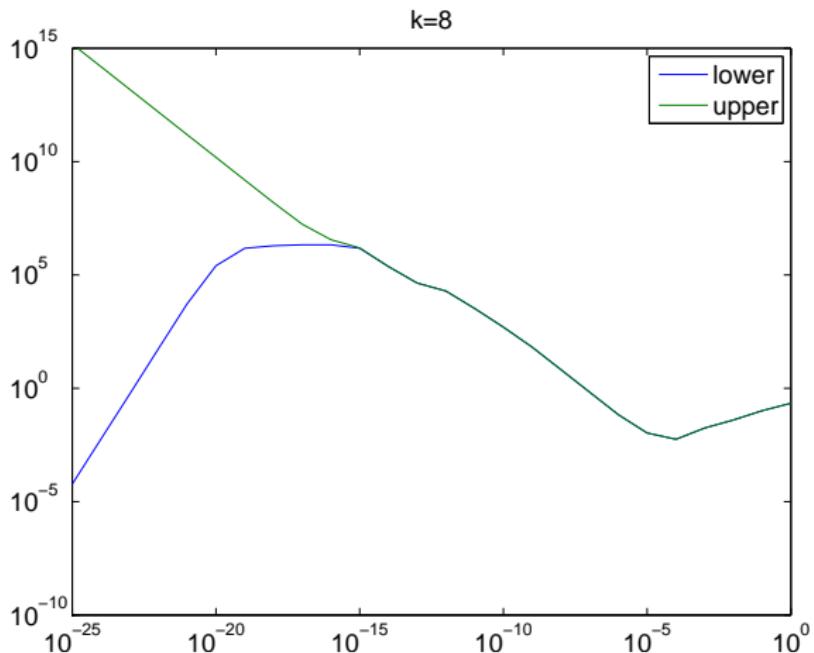
Baart test problem, $m = n = 200$, $\sigma = 10^{-4}$

The algorithm at work: first phase



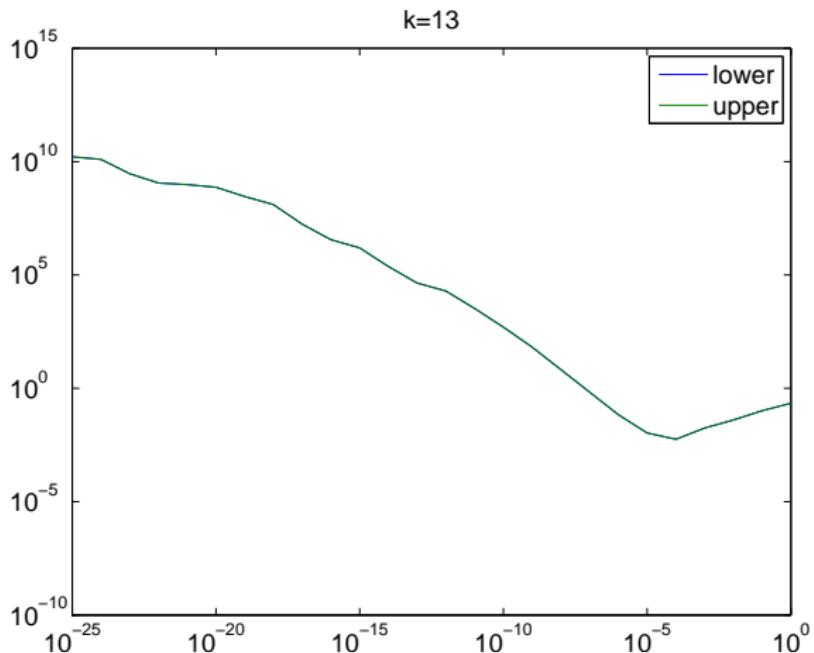
Baart test problem, $m = n = 200, \sigma = 10^{-4}$

The algorithm at work: first phase



Baart test problem, $m = n = 200, \sigma = 10^{-4}$

The algorithm at work: first phase



Baart test problem, $m = n = 200, \sigma = 10^{-4}$

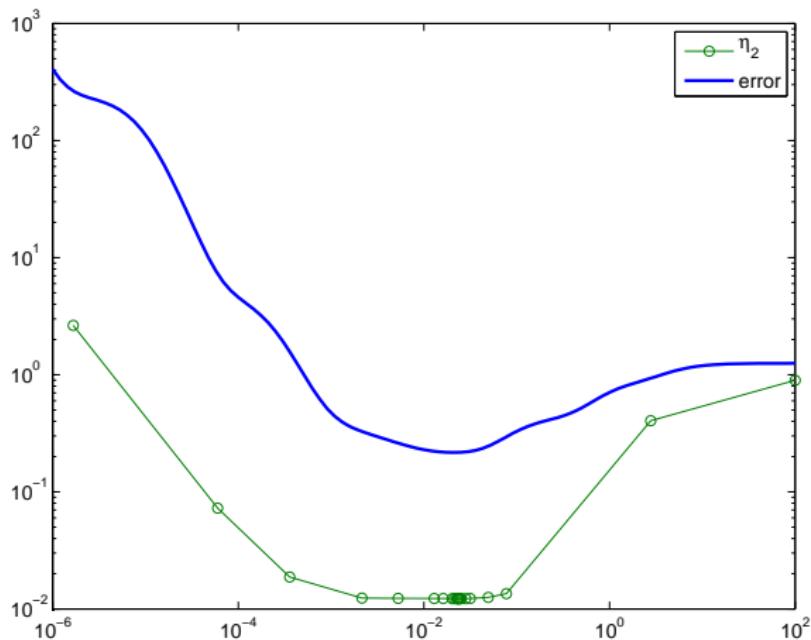
The numerical method: second phase

- Start with the grid \mathcal{C} and the approximate local minimum μ^* determined in phase one.
- Work in log scale: minimize $\eta_\nu(\mu) = \eta_\nu(10^\xi)$.
- Add two points around the minimum by *bisection*.
- Iterate until

$$\min |\xi_{i+1} - \xi_i| < \delta \quad \text{or} \quad \#\{\xi_i\} > N_{\max}.$$

- These computations typically do not require that additional Lanczos bidiagonalization steps be carried out and therefore are inexpensive.

The algorithm at work: second phase



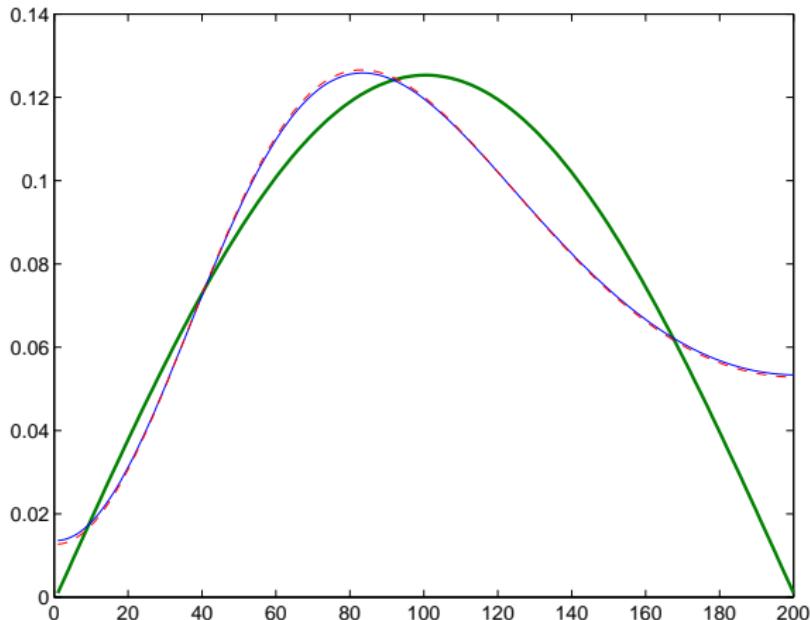
Baart test problem, $m = n = 200$, $\sigma = 10^{-2}$

Computing the solution

Once the parameter μ has been selected, we obtain the Tikhonov regularized solution by computing $\mathbf{x}_\mu = \mathbf{V}_\ell \mathbf{y}_\mu$, where \mathbf{y}_μ is the solution of the least squares problem

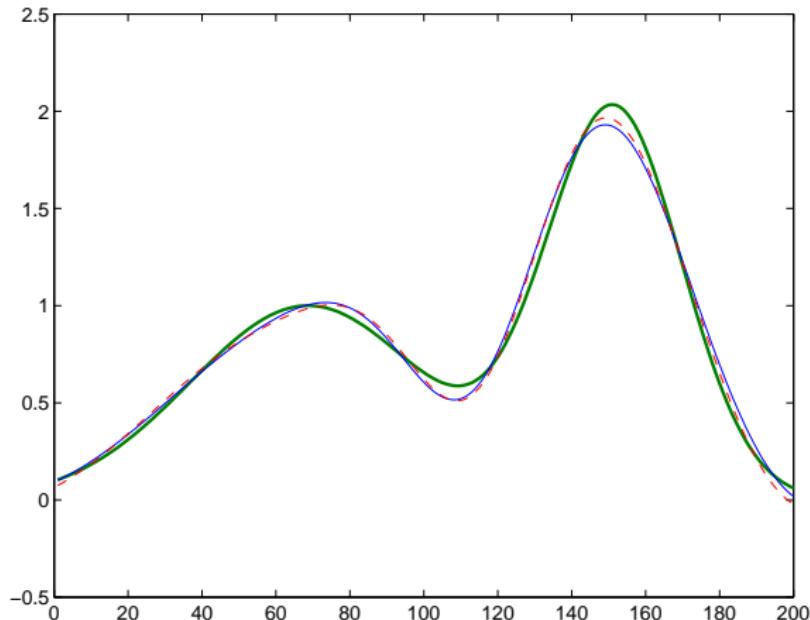
$$\min_{\mathbf{y} \in \mathbb{R}^\ell} \left\| \begin{bmatrix} \bar{\mathcal{C}}_\ell \\ \mu^{1/2} \mathbf{I}_\ell \end{bmatrix} \mathbf{y} - \|\mathbf{b}\| \mathbf{e}_1 \right\|.$$

Numerical examples



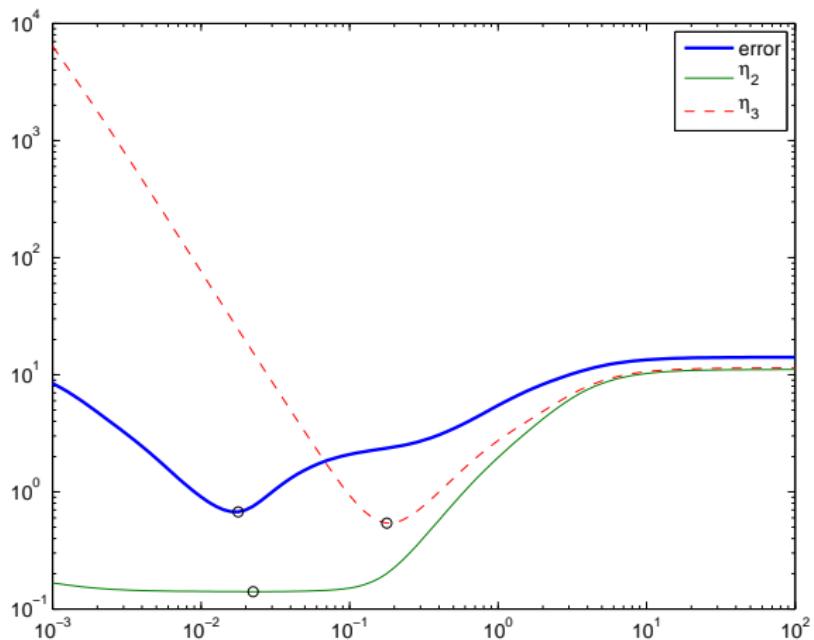
Baart test problem, $m = n = 200$, $\sigma = 10^{-2}$
solution \hat{x} , LBDTIK $x_{\mu,7}$, optimal x_{opt}

Numerical examples



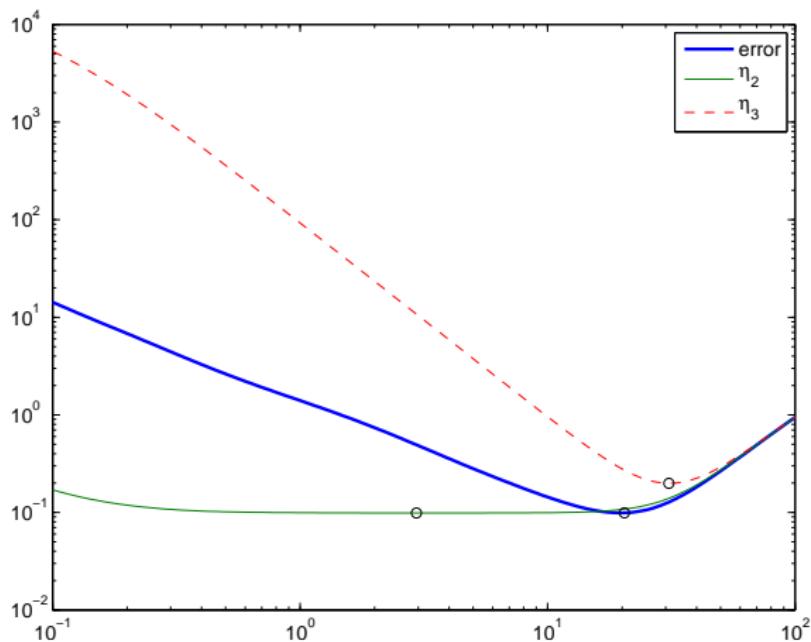
Shaw test problem, $m = n = 200$, $\sigma = 10^{-2}$
solution \hat{x} , LBDTIK $x_{\mu,7}$, optimal x_{opt}

Numerical examples



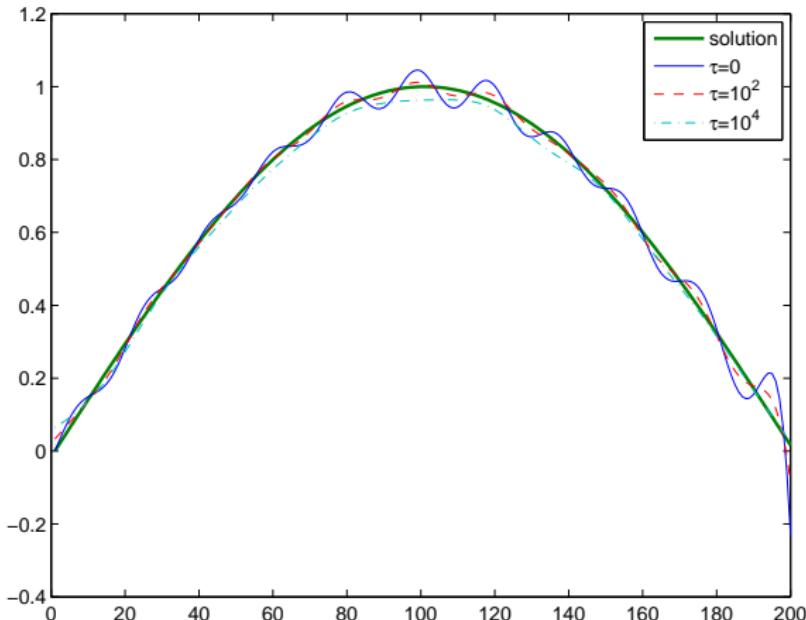
Shaw test problem, $m = n = 200$, $\sigma = 10^{-2}$

Numerical examples



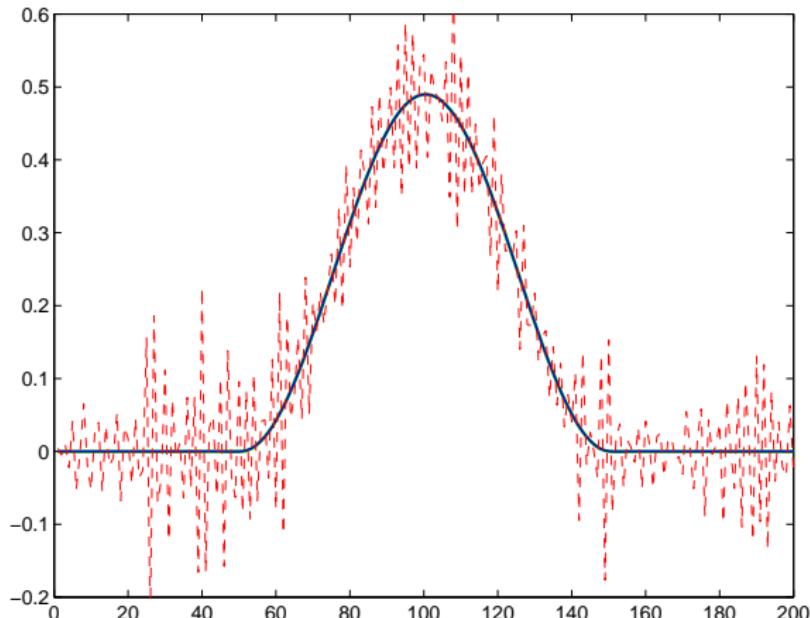
Gauss test problem, $(m, n) = (400, 200)$, $\sigma = 10^{-2}$

Numerical examples



Gauss test problem, $(m, n) = (400, 200)$, $\sigma = 10^{-2}$
Incompatible linear systems with $\tau = \|\hat{\mathbf{b}} - A\hat{\mathbf{x}}\| = 0, 10^2, 10^4$

Numerical examples



Phillips test problem, $m = n = 200$, $\sigma = 10^{-6}$

solution \hat{x} , LBDTIK $x_{\mu,7}$, L-curve x_μ

