Application of Krylov methods to bounding matrix functionals

Mohammed Bellalij

Caterina Fenu

Lothar Reichel

Giuseppe Rodriguez

Hassane Sadok

Dedicated to Mohammed Bellalij

Based on work by Khalide Jbilou, Abderrahim Messaoudi, Gérard Meurant ...

Outline

Bounding the trace of a large implicitly defined matrix

- Application to the solution of ill-posed problems
- Application to network analysis

Example of an ill-posed problem:

Fredholm integral equation of the first kind,

$$\int_0^1 k(s,t)x(t)dt = f(t), \qquad 0 \le s \le 1,$$

with a continuous kernel k.

By the Riemann–Lebesgue lemma, small perturbations in f may correspond to large perturbations in x:

$$\max_{0\le s\le 1} \left| \int_0^1 k(s,t)\cos(2\pi\ell t)dt \right|$$

can be made "tiny" by choosing $|\ell|$ large.

Computed example:

$$\int_0^\pi \exp(-st)x(t)dt = 2\frac{\sinh(s)}{s}, \quad 0 \le s \le \frac{\pi}{2}.$$

Determine solution $x(t) = \sin(t)$.

Discretize integral by Galerkin method using piecewise constant functions. Code baart from Regularization Tools by Hansen.

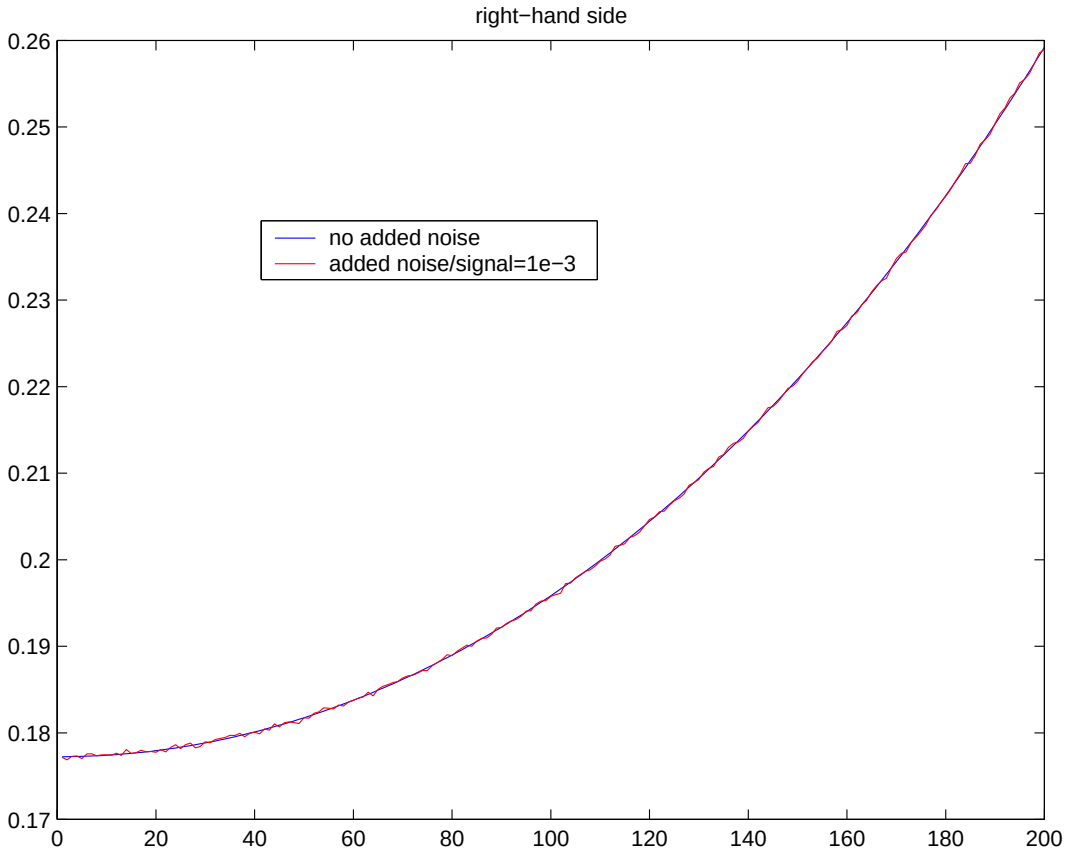
This gives a linear system of equations

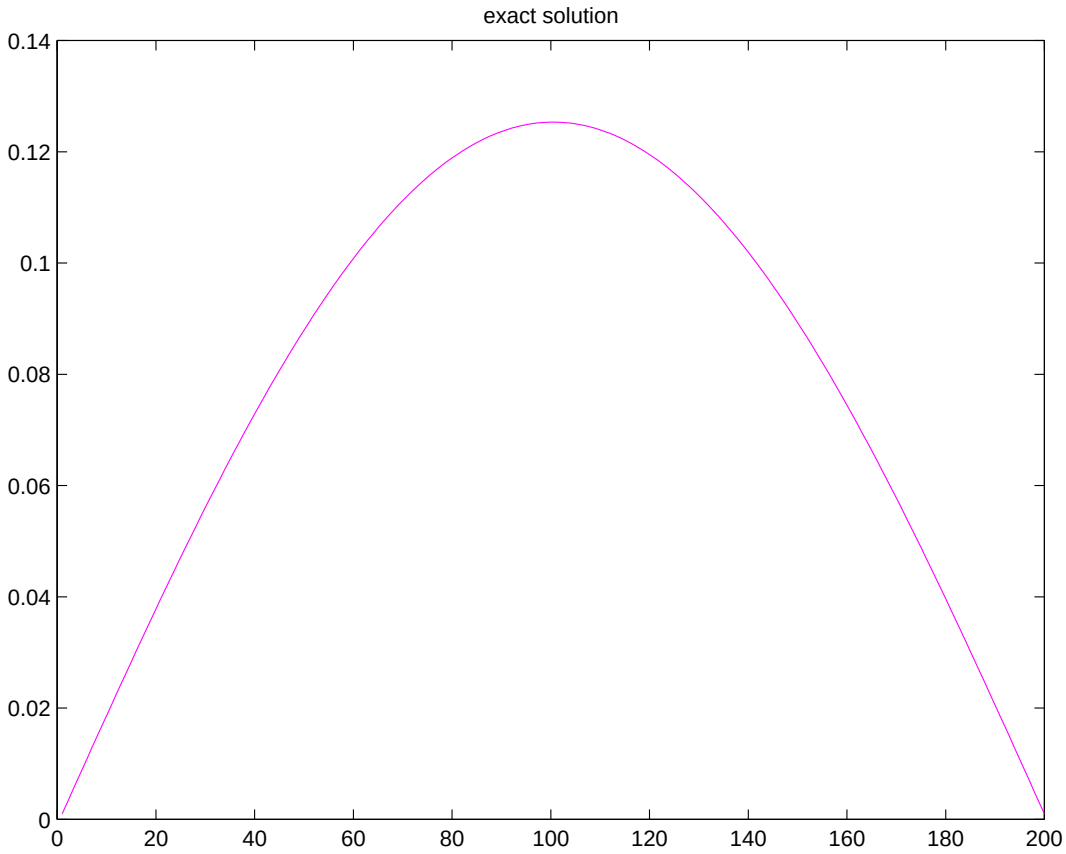
$$Ax = b_{\text{exact}}, \qquad A \in \mathbb{R}^{200\times 200}, \quad b_{\text{exact}} \in \mathbb{R}^{200}.$$

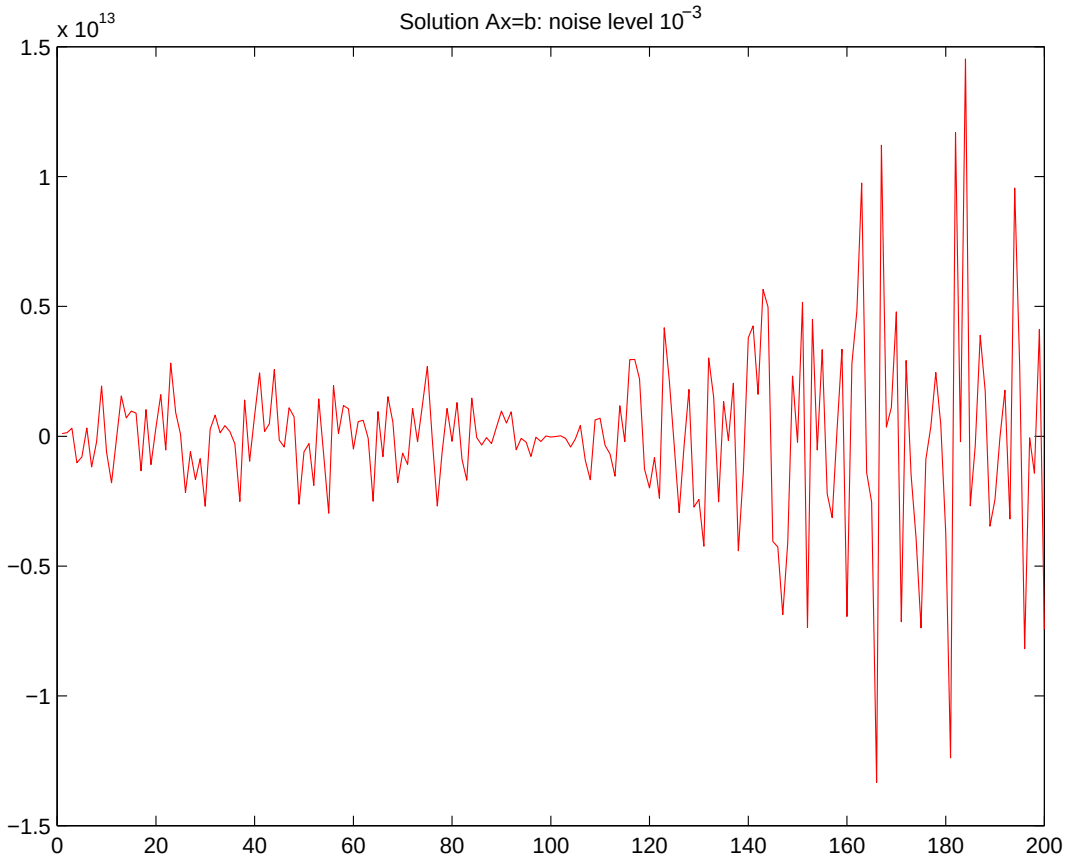
A is numerically singular.

Add "noise vector" e in b that models measurement errors. Let e have e normally distributed entries with mean zero and be scaled to correspond to 0.1% relative error. Thus,

$$b := b_{\text{exact}} + e.$$


right−hand side
no added noise
added noise/signal=1e−3
0.26
0.25
0.24
0.23
0.22
0.21
0.2
0.19
0.18
0.17
0
20
40
60
80
100
120
140
160
180
200


exact solution
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
120
140
160
180
200


Solution Ax=b: noise level 10^-3
x 10^13
1.5
1
0.5
0
−0.5
−1
−1.5
0
20
40
60
80
100
120
140
160
180
200

Tikhonov regularization

Solve the penalized least-squares problem

$$\min_x \{\|Ax - b\|^2 + \mu^2 \|x\|^2\}.$$

Here $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$, $x \in \mathbf{R}^n$, and $\mu > 0$ is a regularization parameter.

- The minimization problem has a unique solution x_μ for any $\mu > 0$.
- It is important to choose an appropriate value of $\mu > 0$. A too small value gives large propagated error in x_μ; a too large value gives an over-smoothed approximate solution.

Choice of the regularization parameter μ

Several ways to determine a suitable μ-value are described in the literature including

- Generalized Cross Validation (GCV),
- the L-curve criterion,
- the discrepancy principle (DP).

DP requires a fairly sharp bound for the error in b to be known. We focus on GCV.

The GCV method

determines the minimizer $\mu > 0$ of the GCV function

$$\mathcal{V}(\mu) := \frac{\|Ax_\mu - b\|^2}{(\text{trace}(I - A(\mu)))^2},$$

where

$$A(\mu) := A(A^T A + \mu^2 I)^{-1} A^T$$

is the influence matrix and

$$x_\mu = (A^T A + \mu^2 I)^{-1} A^T b.$$

Thus,

$$A(\mu)b = Ax_\mu.$$

The GCV function is easy to evaluate when the SVD of A is available. We are interested in the situation when A is too large to make the computation of its SVD attractive.

Let $M \in \mathbf{R}^{m\times m}$ be symmetric, f a function such that $f(M)$ is defined. Let $e_i = [0, \ldots, 0, 1, 0, \ldots, 0]^T$ be "axis vectors" and

$$E_j = [e_{(j-1)k+1}, \ldots, e_{\min\{jk,m\}}], \quad j = 1, \ldots, \widetilde{m},$$

be "block axis vectors" with at most k columns,

$$\widetilde{m} := \left\lfloor \frac{m+k-1}{k} \right\rfloor.$$

We compute upper and lower bounds for

$$\text{trace}(E_j^T f(M) E_j), \qquad j = 1, \ldots, \widetilde{m}$$

with the aid of Gauss quadrature. These bounds give upper and lower bounds for

$$\text{trace}(f(M)) = \sum_{j=1}^{\widetilde{m}} \text{trace}(E_j^T f(M) E_j).$$

Bounding the trace

Define for $\mu > 0$,

$$f_\mu(t) := \frac{\mu^2}{t + \mu^2}, \qquad t \geq 0,$$

Then This follows from

$$\begin{aligned} f_\mu(AA^T) &= \mu^2(AA^T + \mu^2 I)^{-1} \\ &= I - (AA^T + \mu^2 I)^{-1} AA^T \\ &= I - A(A^T A + \mu^2 I)^{-1} A^T \\ &= I - A(\mu). \end{aligned}$$

Introduce the spectral factorization

$$AA^T = Q\Lambda Q^T$$

with Q orthogonal and

$$\Lambda = \mathrm{diag}[\lambda_1, \ldots, \lambda_m], \quad 0 \leq \lambda_1 \leq \ldots \leq \lambda_m.$$

Define for block vectors $U, V \in \mathbf{R}^{m\times k}$ the inner product

$$\langle U, V \rangle := \mathrm{trace}(U^T V)$$

and Frobenius norm

$$\|U\|_F := \langle U, U \rangle^{1/2}.$$

Define the unit block vector

$$\widetilde{W} = [\widetilde{w}_{ij}] = Q^T W / ||W||_F.$$

Then

$$\mathcal{I} f_\mu := \text{trace}(W^T f_\mu(AA^T)W) = \|W\|_F^2 \text{trace}(\widetilde{W}^T f_\mu(\Lambda)\widetilde{W})$$

and

$$e_j^T \widetilde{W}^T f_\mu(\Lambda)\widetilde{W} e_j = \sum_{i=1}^{m} f_\mu(\lambda_i)\widetilde{w}_{ij}^2 = \int_0^\infty f_\mu(\lambda) d\widetilde{w}_j(\lambda),$$

where $\widetilde{w}_j(\lambda)$ is a nondecreasing step function with jumps at the eigenvalues λ_i.

Therefore

$$\mathcal{I}f_\mu = \|W\|_F^2 \sum_{j=1}^{k} \int_0^\infty f_\mu(\lambda) d\widetilde{w}_j(\lambda) = \|W\|_F^2 \int_0^\infty f_\mu(\lambda) d\widetilde{w}(\lambda)$$

with the step function

$$\widetilde{w}(\lambda) := \sum_{j=1}^{k} \widetilde{w}_j(\lambda).$$

Using the fact that $\lambda \to f_\mu(\lambda)$ is totally monotonic (all even order derivatives are positive, all odd order derivatives are negative) will allow us to determine upper and lower bounds for $\mathcal{I}f_\mu$ for $\mu > 0$.

Global Golub–Kahan bidiagonalization (GGKB)

ℓ steps of GGKB applied to A with initial unit block vector U_1 gives the decompositions

$$\begin{aligned} A[V_1, V_2, \ldots, V_\ell] &= [U_1, U_2, \ldots, U_\ell]\widehat{C}_\ell + \sigma_{\ell+1} U_{\ell+1} E_\ell^T, \\ A^T[U_1, U_2, \ldots, U_\ell] &= [V_1, V_2, \ldots, V_\ell]\widehat{C}_\ell^T, \end{aligned}$$

where $U_i \in \mathbf{R}^{m\times k}$, $V_i \in \mathbf{R}^{n\times k}$ and

$$\langle U_i, U_j \rangle = \langle V_i, V_j \rangle = \delta_{ij}.$$

Moreover,

$$\widehat{C}_\ell = C_\ell \otimes I_k,$$

where $\otimes$ stands for Kronecker product, and C_ℓ is a lower bidiagonal matrix,

$$C_\ell = \begin{bmatrix} \rho_1 & & & & \\ \sigma_2 & \rho_2 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \rho_{\ell-1} & \\ & & & \sigma_\ell & \rho_\ell \end{bmatrix} \in \mathbf{R}^{\ell \times \ell}.$$

We also will need

$$C_{\ell+1,\ell} = \begin{bmatrix} C_\ell \\ \sigma_{\ell+1} e_\ell^T \end{bmatrix} \in \mathbf{R}^{(\ell+1)\times\ell}.$$

Combining the GGKB decompositions gives

$$AA^T[U_1, U_2, \ldots, U_\ell] = [U_1, U_2, \ldots, U_\ell]\widehat{T}_\ell + \rho_\ell \sigma_{\ell+1} U_{\ell+1} E_\ell^T,$$

where

$$\widehat{T}_\ell := T_\ell \otimes I_k, \qquad T_\ell := C_\ell C_\ell^T.$$

Theorem: The symmetric tridiagonal matrix T_ℓ defines an ℓ-point Gauss quadrature rule

$$\mathcal{G}_\ell f_\mu = \|W\|_F^2 \, e_1^T f_\mu(T_\ell) e_1$$

associated with the distribution function $\widetilde{w}(\lambda)$, i.e.,

$$\mathcal{G}_\ell p = \mathcal{I} p \qquad \forall p \in \mathbb{P}_{2\ell-1},$$

where $\mathbb{P}_{2\ell-1}$ denotes the set of all polynomials of degree at most $2\ell - 1$.

Remainder formula for Gauss quadrature:

$$\mathcal{I} f_\mu - \mathcal{G}_\ell f_\mu = \frac{f_\mu^{(2\ell)}(\xi)}{(2\ell)!} \int_0^\infty \prod_{j=1}^{\ell} (\lambda - \theta_j)^2 \, d\widetilde{w}(\lambda),$$

where $\xi \in (\lambda_1, \lambda_m)$ and the θ_j are Gaussian nodes. They satisfy $\lambda_1 < \theta_j < \lambda_m$ for all j. It follows that

$$\mathcal{G}_{\ell-1} f_\mu < \mathcal{G}_\ell f_\mu < \mathcal{I} f_\mu.$$

For discussions on the relation between the standard Lanczos process and Gauss quadrature; see

G. H. Golub and G. Meurant: Matrices, Moments and Quadrature with Applications, Princeton, 2010.

An $(\ell+1)$-point Gauss–Radau rule associated with the distribution function $\widetilde{w}$ and a fixed node $\zeta = 0$ can be expressed as

$$\mathcal{R}_{\ell+1,0} f_\mu = \|W\|_F^2 e_1^T f_\mu(T_{\ell+1,0}) e_1,$$

where

$$T_{\ell+1,0} = C_{\ell+1,\ell} C_{\ell+1,\ell}^T$$

and

$$\mathcal{R}_{\ell+1,0} p = \mathcal{I} p \qquad \forall p \in \mathbb{P}_{2\ell}.$$

Note that $T_{\ell+1,0}$ is singular.

Remainder formula for Gauss–Radau quadrature:

$$\mathcal{I}f_\mu - \mathcal{R}_{\ell+1,0}f_\mu = \frac{f_\mu^{(2\ell+1)}(\xi_0)}{(2\ell+1)!}\int_0^\infty \lambda \prod_{j=1}^{\ell}(\lambda-\theta_{j,0})^2\, d\widetilde{w}(\lambda).$$

The $\theta_{j,0}$ are "free" Gauss-Radau nodes. Odd derivatives $f_\mu^{(2\ell+1)}(t)$ are negative for $t \geq 0$. One can show that

$$\mathcal{I}f_\mu < \mathcal{R}_{\ell+1,0}f_\mu < \mathcal{R}_{\ell,0}f_\mu.$$

Bounding the numerator of the GCV function

Let $x_\mu = (A^T A + \mu^2 I)^{-1} A^T b$ be the Tikhonov solution for $\mu > 0$, and define

$$g_\mu(t) := \frac{\mu^2}{(t+\mu^2)^2}, \qquad t \geq 0.$$

Then

$$\|Ax_\mu - b\|^2 = \mu^2 b^T g_\mu(AA^T) b.$$

Substituting the spectral factorization of AA^T into the rhs gives

$$\mu^2 b^T g_\mu(AA^T) b = \mu^2 \|b\|^2 \int_0^\infty g_\mu(\lambda) d\nu(\lambda).$$

Application of ℓ steps of (standard) Golub–Kahan bidiagonalization to A with initial vector b gives the decompositions

$$\begin{aligned} A[v_1, v_2, \ldots, v_\ell] &= [u_1, u_2, \ldots, u_\ell]B_\ell + \beta_{\ell+1}u_{\ell+1}e_\ell^T, \\ A^T[u_1, u_2, \ldots, u_\ell] &= [v_1, v_2, \ldots, v_\ell]B_\ell^T, \end{aligned}$$

where

$$u_i^T u_j = \delta_{ij}, \qquad v_i^T v_j = \delta_{ij},$$

and $B_\ell \in \mathbf{R}^{\ell\times\ell}$ is lower bidiagonal.

Let $\check{T}_\ell := B_\ell B_\ell^T$. Then

$$\check{\mathcal{G}}_\ell g_\mu := \mu^2 \|b\|^2 \, e_1^T g_\mu(\breve{T}_\ell) e_1$$

is an ℓ-point Gauss quadrature rule associated with the measure $d\nu$, i.e.,

$$\check{\mathcal{G}}_\ell p = \mu^2 \|b\|^2 \int_0^\infty p(\lambda) d\nu(\lambda), \qquad \forall p \in \mathbb{P}_{2\ell-1}.$$

Moreover,

$$\check{\mathcal{G}}_{\ell-1} g_\mu < \check{\mathcal{G}}_\ell g_\mu < \|Ax_\mu - b\|^2.$$

One obtains a $(\ell+1)$-point Gauss–Radau rule with a fixed node at zero by appending a row to B_ℓ,

$$\check{\mathcal{R}}_{\ell+1,0} := \mu^2 \|b\|^2 \, e_1^T g_\mu(\breve{B}_{\ell+1,\ell}\breve{B}_{\ell+1,\ell}^T) e_1, \quad \breve{B}_{\ell+1,\ell} = \begin{bmatrix} \breve{B}_\ell \\ \beta_{\ell+1} e_\ell^T \end{bmatrix}.$$

Then

$$\check{\mathcal{R}}_{\ell+1,0} p = \mu^2 \|b\|^2 \int_0^\infty p(\lambda) d\nu(\lambda) \qquad p \in \mathbb{P}_{2\ell}$$

and

$$\|Ax_\mu - b\|^2 < \check{\mathcal{R}}_{\ell+1} g_\mu < \check{\mathcal{R}}_\ell g_\mu.$$

Computed examples

We have described how the connection between Golub–Kahan bidiagonalization and Gauss quadrature can be exploited to determine upper and lower bounds for the numerator and denominator for the GCV function

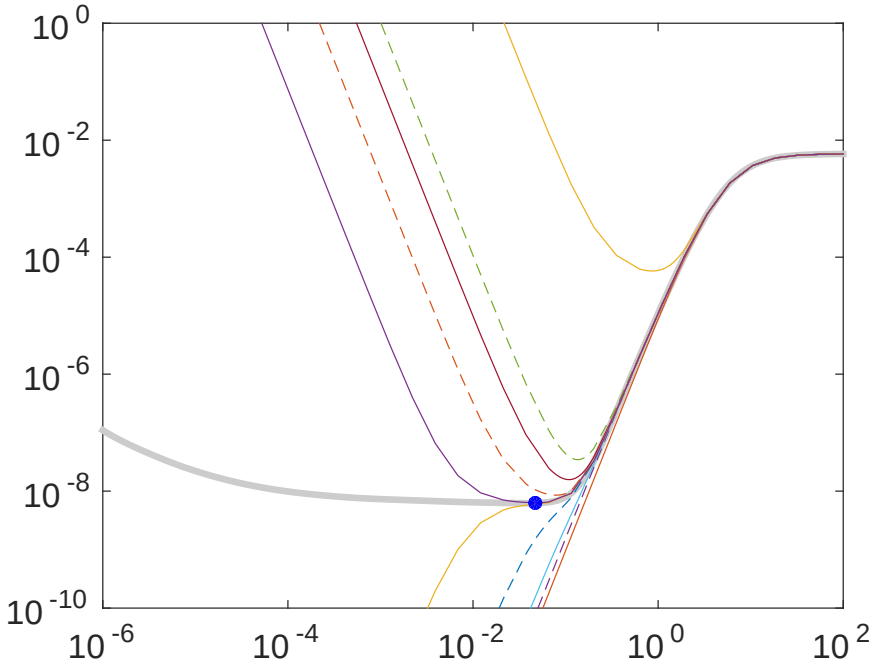
$$\mathcal{V}(\mu) = \frac{\|Ax_\mu - b\|^2}{(\mathrm{trace}(I - A(\mu)))^2}.$$

We therefore can compute upper and lower bounds for $\mathcal{V}(\mu)$.

Example 1: Consider the test problem phillips from Regularization Tools,

$$Ax = b, \qquad A \in \mathbf{R}^{200\times 200}, \quad x, b \in \mathbf{R}^{200}.$$

0.1% relative error in b.

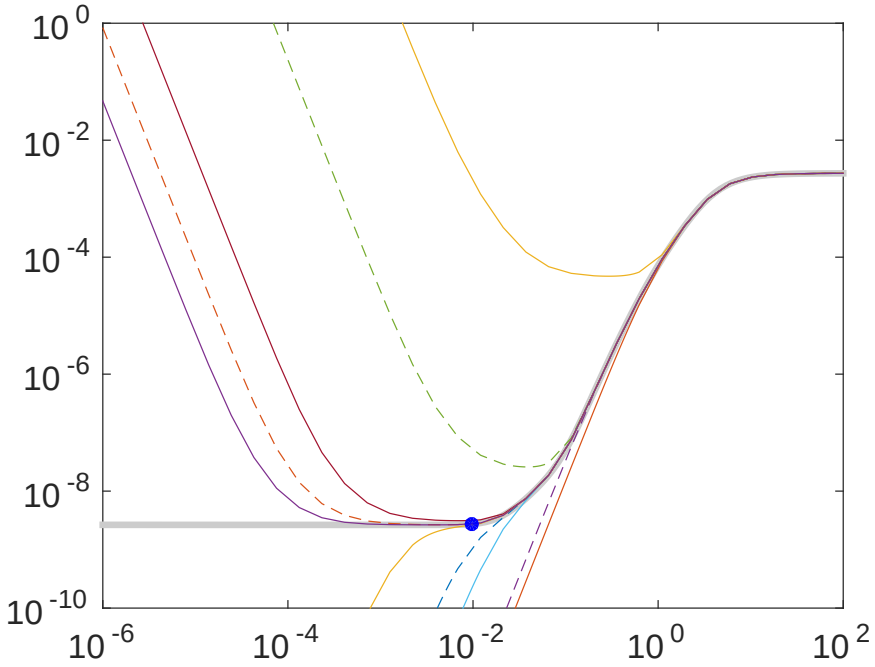


Upper and lower bounds for $\mathcal{V}(\mu)$ for $\ell = 2, 4, \ldots, 10$.

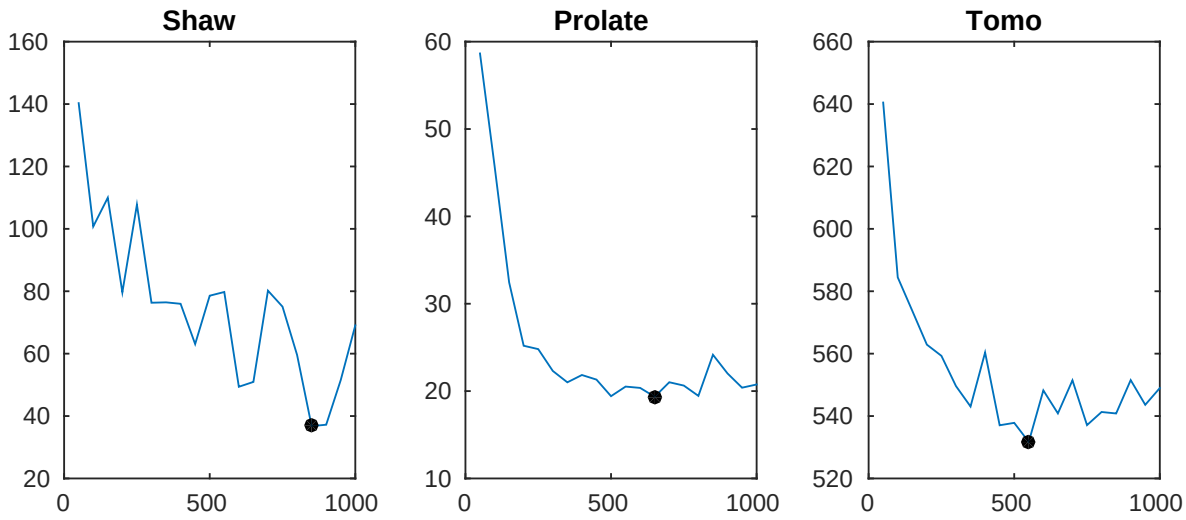
Example 2: Consider the test problem shaw from Regularization Tools,

$$Ax = b, \qquad A \in \mathbf{R}^{2000\times 2000}, \quad x, b \in \mathbf{R}^{2000}.$$

0.1% relative error in b.



Upper and lower bounds for $\mathcal{V}(\mu)$ for $\ell = 2, 4, \ldots, 10$.



Example 3: Timings as a function of block size. $A \in \mathbf{R}^{4096 \times 4096}$.

Example 4: Problems from Regularization Tools with $A \in \mathbf{R}^{2000\times 2000}$ or $A \in \mathbf{R}^{4000\times 2000}$. Block size $k = 500$. Relative error in b: $0.1, 0.01, 0.001$. 10 runs per noise level gives 60 runs per matrix-type.

`gcv_lanczos` implementation by Golub and von Matt using Hutchinson's stochastic trace estimator: Compute upper and lower bounds of

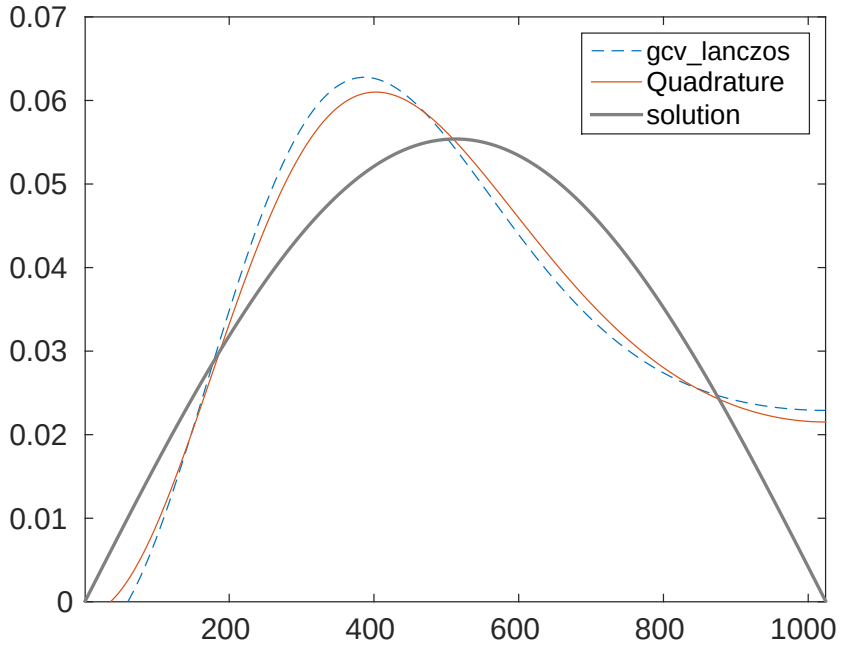
$$z^T(AA^T + \mu^2 I)^{-1} z,$$

z random vector with entries ± 1 with equal probability.

F_c = number of times error is c times larger than best Tikhonov. Block size $k = 500$.

	`gcv_lanczos`			Quadrature		
matrix	F_5	F_{10}	time	F_5	F_{10}	time
Baart	27	22	7.0e-01	0	0	2.0e+01
Deriv2	0	0	7.4e-01	0	0	2.9e+01
Foxgood	33	30	4.9e-01	3	0	1.7e+01
Gravity	9	7	6.2e-01	1	0	3.4e+01
Heat(1)	0	0	1.4	0	0	4.1e+01
Phillips	4	1	4.0e-01	0	0	3.3e+01
Shaw	25	21	9.1e-01	0	0	2.1e+01
Wing	19	18	6.0	0	0	1.8e+01

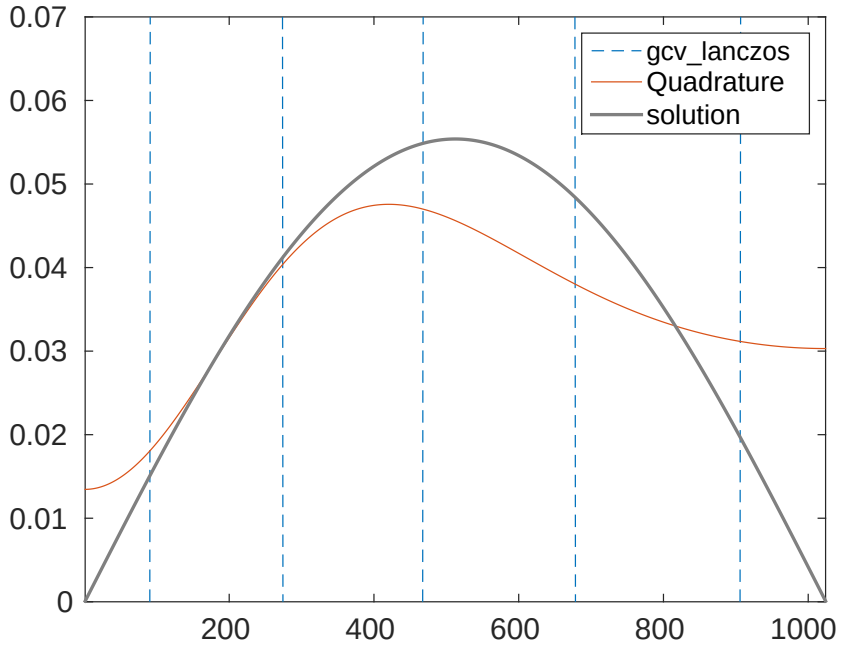
Example 5: Baart test problem $A \in \mathbf{R}^{1024\times 1024}$. Block size $k = 512$. Relative error in b: 0.1.



Quadrature: $\mu = 2.3 \cdot 10^{-2}$, rel. error$=2.1 \cdot 10^{-1}$

gcv_lanczos: $\mu = 4.0 \cdot 10^{-1}$, rel. error$=2.5 \cdot 10^{-1}$

Example 5: Baart test problem $A \in \mathbf{R}^{1024\times 1024}$. Block size $k = 512$. Relative error in b: 0.1.



Quadrature: $\mu = 4.0 \cdot 10^{-2}$, rel. error$=2.5 \cdot 10^{-1}$

gcv_lanczos: $\mu = 5.1 \cdot 10^{-6}$, rel. error$=1.3 \cdot 10^{2}$

Application to network analysis:

$A \in \mathbf{R}^{n \times n}$ a symmetric matrix, $W \in \mathbf{R}^{n \times k}$ initial block vector for the global block Lanczos method, block size $k \ll n$. This method uses the inner product

$$\langle U, V \rangle = \frac{1}{k} \text{trace}(U^T V), \qquad U, V \in \mathbf{R}^{n \times k}$$

and norm

$$\|V\| = \langle V, V \rangle^{1/2}.$$

The global block Lanczos process

ℓ steps determine an ON basis $\{V_k\}_{k=1}^{\ell+1}$ for the block Krylov subspace $\mathcal{K}_{\ell+1}(A, W) := \text{span}\{W, AW, \ldots, A^{\ell}W\}$.

Algorithm:

$V_0 := 0,\ \beta_1 := \|W\|;\ V_1 := W/\beta_1;$

for $j = 1, 2, \ldots, \ell$ **do**

$\widetilde{V} = AV_j - \beta_j V_{j-1};\ \alpha_j = \langle V_j, \widetilde{V} \rangle;$

$\widetilde{V} = \widetilde{V} - \alpha_j V_j;$

$\beta_{j+1} = \|\widetilde{V}\|;\ V_{j+1} = \widetilde{V}/\beta_{j+1};$

end

The global Lanczos process gives the decomposition

$$A[V_1, \ldots, V_\ell] = [V_1, \ldots, V_\ell]\widehat{T}_\ell + \gamma_{\ell+1} V_{\ell+1} E_\ell^T,$$

where $V_j \in \mathbf{R}^{n\times k}$ with

$$\langle V_i, V_j \rangle = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

and

$$\widehat{T}_\ell = T_\ell \otimes I_k.$$

Application: Let

$f(t) = \exp(t), \quad A$ adjacency matrix, $\quad W = E_1 = [e_1, e_2, \ldots, e_k]$.

Pairs of Gauss and Gauss-Radau rules provide lower and upper bounds for

$$\text{trace}(E_1^T \exp(A) E_1).$$

Repeat for

$$W = E_j, \quad j = 2, 3, \ldots, n/k.$$

Gives lower and upper bounds for

$$\text{trace}(\exp(A)) = \text{Estrada index for graph.}$$

Table 1: Execution times for computing the Estrada index.

		Glob. Lanczos		Scal. Lanczos	`expm`
matrix	nodes	time	k	time	time
Email	1133	3.45e-01	80	3.54e+00	1.18e+01
Yeast	2114	4.73e-01	60	3.20e+00	1.01e+01
Power	4941	1.89e+00	40	1.28e+01	2.14e+01
Internet	22963	1.22e+02	8	3.30e+02	-
Collaborations	40421	4.80e+02	40	1.25e+03	-
Facebook	63731	2.64e+03	60	8.82e+03	-

Table 2: Number of iterations and matrix-vector product evaluations (MVPs)

			Glob. Lanczos		Scal. Lanczos
matrix	nodes	k	iterations	MVPs	MVPs
Email	1133	80	130	9730	9745
Yeast	2114	60	213	12596	8652
Power	4941	40	618	24644	22209
Internet	22963	8	28776	230158	226433
Collaborations	40421	40	10115	404372	416648
Facebook	63731	60	12075	723912	758618

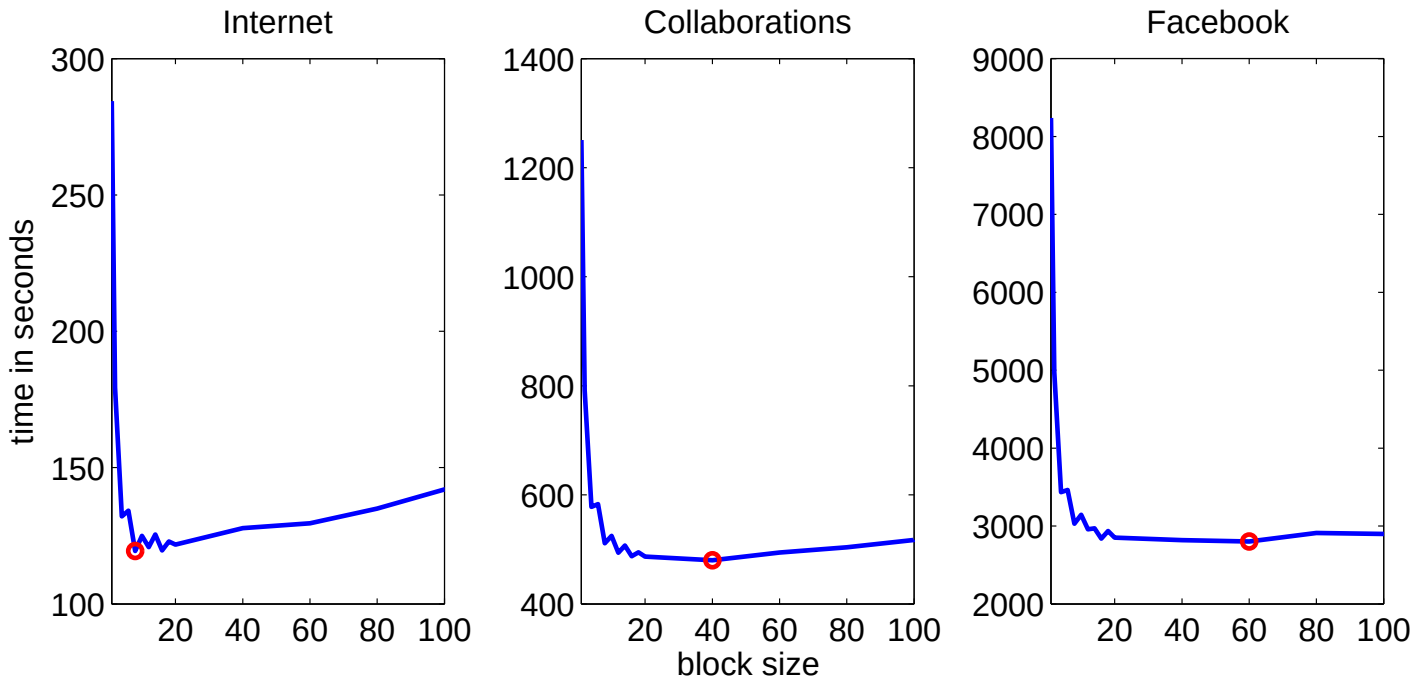


Figure 1: Execution times to compute the Estrada index for the three largest networks, versus the block size k.

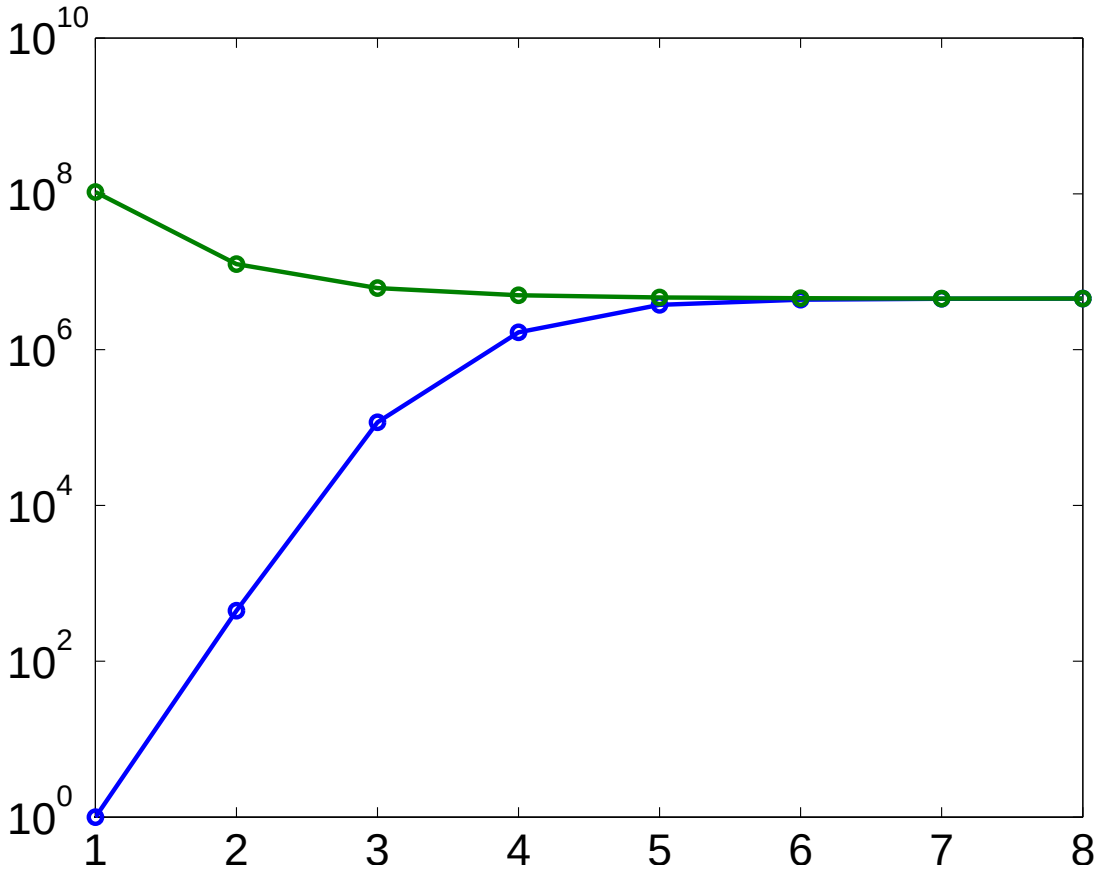


Figure 2: Upper and lower bounds for the Estrada index versus number of iterations for Email network.

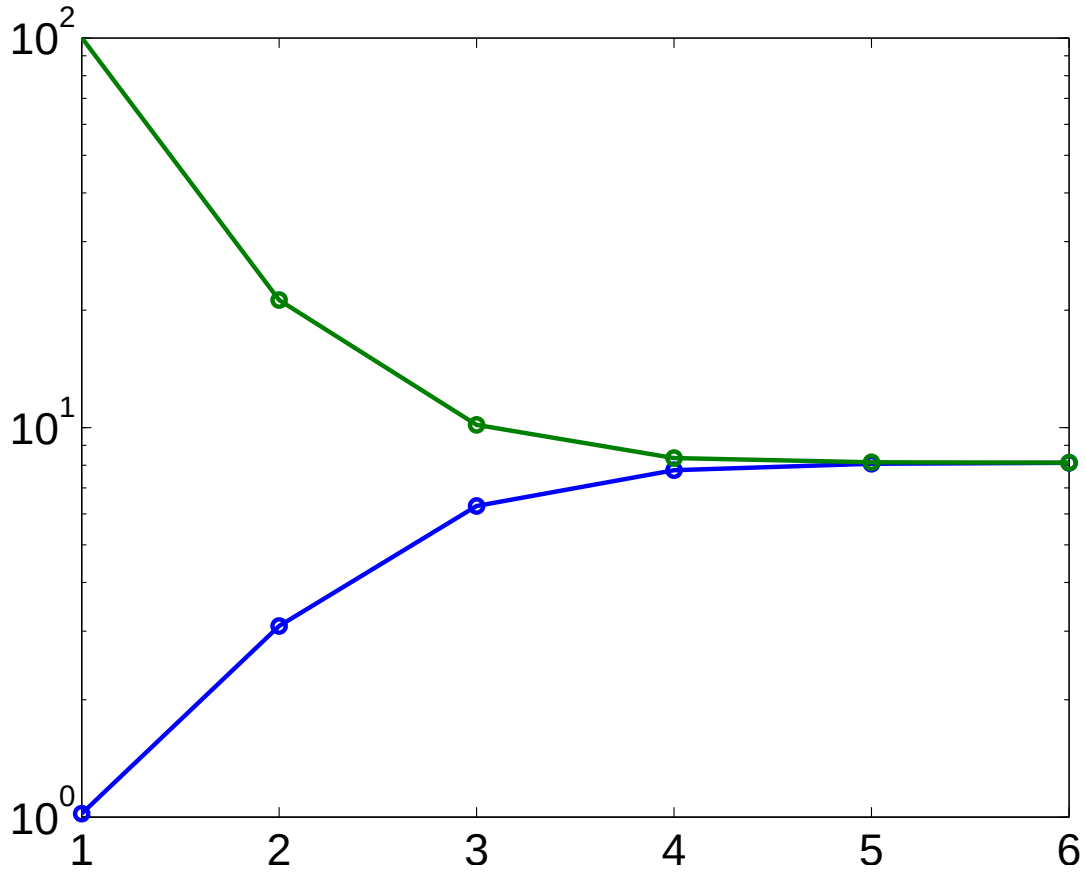


Figure 3: Upper and lower bounds for the Estrada index versus number of iterations for Yeast network.

Merci