Thick-Restart Lanczos method for Symmetric Eigenvalue Problems

Kesheng Wu and Horst Simon
NERSC
Lawrence Berkeley National Laboratory
{KWu, HDSimon}@lbl.gov

March 31, 1998

URL: http://www.nersc.gov/research/SCG/wu

Outline

- Basic Lanczos algorithm
- Thick-restart Lanczos algorithm
- Loss of orthogonality & recovery
- Actual loss of orthogonality
- Accuracies of solutions

Basic Lanczos algorithm

ALGORITHM **1.** *The Lanczos iterations starting with* r_0. *Let* $\beta_0 = \|r_0\|$, *and* $q_0 = 0$.
For $i = 1, 2, \ldots,$

(a) $q_i = r_{i-1}/\|r_{i-1}\|$,

(b) $p = Aq_i$,

(c) $\alpha_i = q_i^T p$,

(d) $r_i = p - \alpha_i q_i - \beta_{i-1} q_{i-1}$,

(e) $\beta_i = \|r_i\|$.

ALGORITHM **2.** *The Arnoldi iterations starting with* r_0.

For $i = 1, 2, \ldots$

(a) $q_i = r_{i-1}/\|r_{i-1}\|$,

(b) $p = Aq_i$,

(c) $h_{j,i} = q_j^T p, j = 1, \ldots, i$,

(d) $r_i = p - \sum_{j=1}^{i} h_{j,i} q_j$,

(e) $h_{i,i+1} = \|r_i\|$.

Basic Lanczos algorithm

Major issues to address

- $\surd$ loss of orthogonality
- $\surd$ storage requirement
- $\surd$ arithmetic operation count
- – convergence rate
- – convergence test
- – parallel implementation
- – efficient application of the operator

Basic Lanczos algorithm

Available strategies to deal with loss of orthogonality

- No re-orthogonalization, identify spurious eigenpairs (Cullumn, Willoughby)
- full re-orthogonalization — always orthogonalize against all basis vectors
- Selective re-orthogonalization — only orthogonalize against converged eigenvectors (Scott, Parlett)
- Partial re-orthogonalization — ω-recurrence (Simon, Parlett)
- Combination of selective and partial re-orthogonalization (BLZPACK Marques)

Basic Lanczos algorithm

Restarting strategies — controlling maximum memory usage

- restarting with one vector
- restarting with a linear combination of several vectors
- implicit-restarting scheme of implicitly restarted Arnoldi (Sorensen)
- thick-restart scheme of dynamic thick-restart Davidson method (Stathopoulos, Saad, Wu)

Thick-restart Lanczos algorithm

Why restarted Lanczos

- successful new restarting schemes
- controlling the maximum amount of memory used
- the Lanczos algorithm uses less arithmetic operations

Thick-restart Lanczos algorithm

ALGORITHM **3.** *Starting with k vectors s.t.* $Aq_i = \alpha_i q_i + \beta_i q_{k+1}$, $i = 1, \ldots, k$.

1 Initialization.

(a) $q_{k+1} = r_k / \|r_k\|$,

(b) $p = Aq_{k+1}$,

(c) $\alpha_{k+1} = q_{k+1}^T p$,

(d) $r_{k+1} = p - \alpha_{k+1} q_{k+1} - \sum_{i=1}^{k} \beta_i q_i$,

(e) $\beta_{k+1} = \|r_{k+1}\|$,

2 Iterate. *For* $i = k+2, \ldots,$

(a) $q_i = r_{i-1} / \beta_{i-1}$,

(b) $p = Aq_i$,

(c) $\alpha_i = q_i^T p$,

(d) $r_i = p - \alpha_i q_i - \beta_{i-1} q_{i-1}$,

(e) $\beta_i = \|r_i\|$.

Thick-restart Lanczos algorithm

Observations about the restarted Lanczos algorithm

- given m, maximum memory usage $n \times (m+1) + \mathrm{O}(m^2)$
- the vectors r_i and p can be stored in the same space
- aside from the initialization step, each step is the same as the original Lanczos algorithm
- mathematically equivalent to implicitly restarted Arnoldi method with exact shifts
- the dynamic restart-scheme for the dynamic thick-restart Davidson method can be used

Loss of orthogonality

- Adopt the ω-recurrence for the restarted Lanczos algorithm.
- Before computing β_i, i.e., after step (d) and before step (e),

  Global re-orthogonalization: if the norm of $(\omega_{i+1,1}, \ldots, \omega_{i+1,i})$ is large, say, greater then $\sqrt{\epsilon}$, use Gram-Schmidt procedure to orthogonalize r_i against all current basis vectors,

  Local re-orthogonalization: else, orthogonalize r_i against q_i (q_{i-1}).

Global re-orthogonalization

Is Gram-Schmidt process effective when Q is semi-orthogonal?

YES! $\Longrightarrow$ Repeating Gram-Schmidt process on an vector $z_{(0)}$, let

$$z_{(i+1)} \equiv (I - QQ^T)z_{(i)}, \quad w_z^{(i)} \equiv Q^T z_{(i)}.$$

- $w_z^{(i)} = (I - QQ^T)w_z^{(i-1)} = (I - QQ^T)^i Q^T z_{(0)}$
- $i \to \infty$ and $\|I - QQ^T\| < 1, \quad w_z^{(i)} \to z_\perp \equiv (I - NN^T)z_{(0)}$
- full orthogonality, $\|I - QQ^T\| \sim \mathrm{O}(\epsilon)$
- semi-orthogonality, $\|I - QQ^T\| \sim \mathrm{O}(\sqrt{\epsilon})$

Global re-orthogonalization

Effects of global re-orthogonalization

- √ r_i is orthogonal to Q_i
- √ norm of r_i is accurate
- √ clean starting point for future iterations
- – need to orthogonalize q_i as well

Local re-orthogonalization

To decrease the error in α_i	To decrease the error in β_i
	extended local re-orthogonalization (Parlett, 1992)
√ $r_i = r_i - q_i q_i^T r_i$	√ $r_i = r_i - q_i q_i^T r_i - q_{i-1} q_{i-1}^T r_i$
√ once is enough	√ once is enough
√ norm of r_i not affected	√ norm of r_i not affected

full versus partial

maintain full orthogonality

- no ω-recurrence
- 2 Gram-Schmidt is enough
- more likely to invoke GS
- accurate solutions

maintain semi-orthogonality

- ω-recurrence
- may need 4 Gram-Schmidt
- less likely to invoke GS
- Ritz values are accurate

Loss of orthogonality

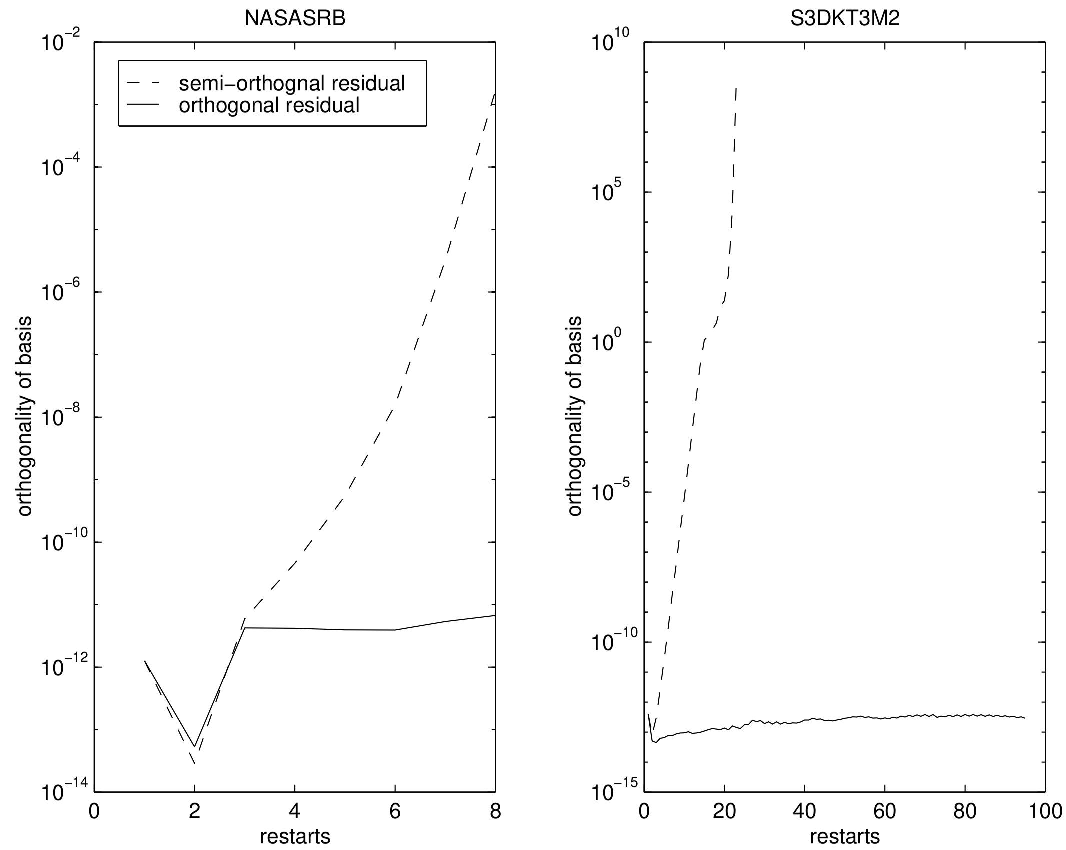


Figure 1: The orthogonality level of $(\|Q^T Q - I\|_F)$.

Accuracies of solutions

name	N	NNZ	$\|A\|$	description
NASASRB	54870	2677324	2.6E9	shuttle rocket booster
S3DKT3M2	90449	3753461	8798	cylindrical shell
S3DKQ4M2	90449	4820891	4602	cylindrical shell

Table 1: Information about the test matrices.

Accuracies of solutions

Two measures of errors

- diff between Rayleigh quotient and Ritz value, δ_λ
- diff between the estimated residual norm and the actual residual norm, δ_r

$$\boxed{\delta_\lambda \sim \mathrm{O}(\epsilon\|A\|)}$$

$$\boxed{\delta_r \sim \mathrm{O}(\epsilon\|A\|),\ \text{full};\quad \delta_r \sim \mathrm{O}(\sqrt{\epsilon}\|A\|),\ \text{semi}}$$

Accuracies of solutions

Methods tested

TRLan PRO thick-restart Lanczos method with partial re-orthogonalization

TRLan PRO-c thick-restart Lanczos method with partial re-orthogonalization and correction to compute orthogonal basis at restart

TRLan full thick-restart Lanczos method that maintains full orthogonality

PARPACK the parallel version of ARPACK (Sorensen, etc.)

Accuracies of solutions

NASASRB, 5 largest eigenvalues, basis size 20

	TRLan PRO	TRLan PRO-c	TRLan full	PARPACK
MATVEC	90	90	90	157
restarts	8	8	8	11
δ_λ	4E-6	4E-6	2E-6	3E-6
δ_r	8E-6	7E-6	2E-6	2E-6

NASASRB, 5 largest eigenvalues, basis size 10

	TRLan PRO	TRLan PRO-c	TRLan full	PARPACK
MATVEC	185	185	185	184
restarts	46	46	46	39
δ_λ	7E-6	9E-6	2E-5	6E-6
δ_r	1E-5	2E-6	2E-6	3E-6

Accuracies of solutions

S3DKT3M2, 5 largest eigenvalues, basis size 10

	TRLan PRO	TRLan PRO-c	TRLan full	PARPACK
MATVEC	2269	2269	2269	4459
restarts	465	465	465	1310
δ_λ	1E-10	1E-10	4E-11	2E-10
δ_r	5E-11	5E-11	2E-10	7E-11

S3DKQ4M2, 5 largest eigenvalues, basis size 10

	TRLan PRO	TRLan PRO-c	TRLan full	PARPACK
MATVEC	5119	5113	5119	3646
restarts	1516	1514	1516	1516
δ_λ	5E-10	4E-10	4E-10	3E-11
δ_r	6E-10	4E-10	3E-10	4E-11

Accuracies of solutions

NASASRB, 5 smallest eigenvalues, basis size 1000

	TRLan PRO	TRLan full	PARPACK
MATVEC	50471	50467	46761
restarts	51	51	61
δ_λ	9E-7	2E-6	5E-7
δ_r	7E-4	3E-6	4E-6

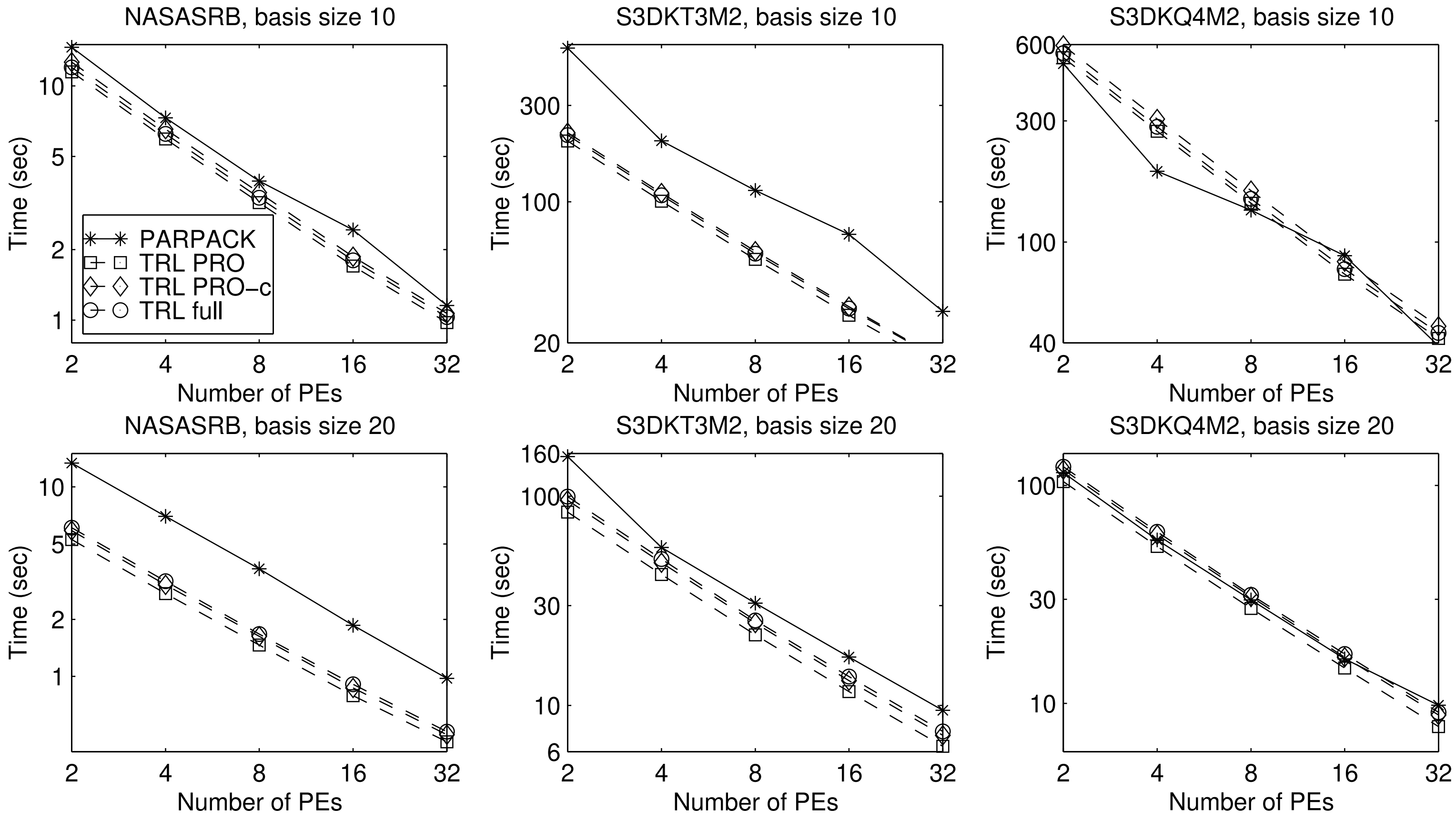
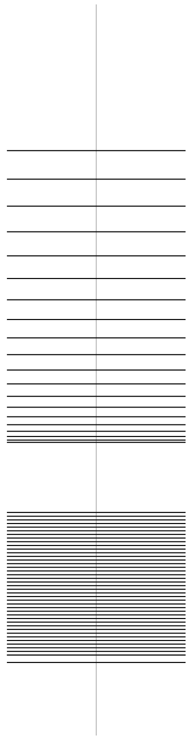


Figure 2: Time spent to solve the three test problems on Cray T3E900.

In a real application

`escan` (Wang and Zunger, NREL)

- semi-conductor quantum dot model
- represent solution as Fourier components
- specialized FFT for T3E (Canning)
- rectangular real space grid
- want interior eigenvalues of
  $H = -\nabla^2 + V$
- compute the smallest eigenvalues
  $(H - E_{ref} I)^2$



In a real application

Comparing TRLan with CG in `escan`
time in seconds, on Cray T3E at NERSC

# eig	valence CG	valence TRLan	conduction CG	conduction TRLan
10	172.4	63.5	444.7	110.7
100	2236.3	449.6	2584.1	709.7

Table 2: interior eigenvalues, InGaP, 512 atoms, 6603 G, 48 $\times$ 48 $\times$ 48 grid, 8 PE.

In a real application

# eig	CG	TRLan	PLANSO
1	2.1	2.0	1.1
5	20.1	7.0	6.7
10	43.3	11.4	11.8
20	60.1	11.2	12.5
50	237.8	29.7	70.4
100	618.0	52.7	138.5

Table 3: smallest eigenvalues, InGaP, 512 atoms, 6603 G, $48 \times 48 \times 48$ grid, 8 PE.

In a real application

# eig	CG	TRLan	PLANSO
1	115.0	72.0	59.0
10	380.0	164.5	142.4
20	704.3	184.8	172.9
100	4608.8	612.0	—

Table 4: smallest eigenvalues, InGaAs, 9K atoms, 137,919 G, $240 \times 36 \times 320$ grid, 32 PE.

Summary

When to use thick-restart Lanczos method

- $\surd$ symmetric standard eigenvalue problem
- $\surd$ only matrix-vector multiplication
- $\surd$ non-restarted Lanczos takes too many steps
- $\surd$ wants large number of eigenvalues
- $\surd$ no initial guess

Alternative

- ARPACK
- Arnoldi
- Davidson
- ...