Making the Lanczos method work for electronic structure calculations

Kesheng Wu

Andrew Canning

Horst D. Simon

NERSC, Lawrence Berkeley National Laboratory

{kwu, acanning, hdsimon}@lbl.gov

Outline

1. Background: electronic structure calculation, Lanczos method
2. Thick-restart Lanczos algorithm
3. How to restart
4. Performance characteristics

Electronic structure calculations

1. Schrodinger Equation: $H\Psi = E\Psi$
2. Density Functional Theorem(DFT): Hohenberg-Kohn (1964)
3. Kohn-Sham equation + Local Density Approximation + pseudopotential,...
4. <u>many discretization schemes lead to matrix eigenvalue problems</u>

Characteristics of the eigenvalue problems

- Large matrices
- Fast matrix-vector multiplication but matrix may not be stored
- Many eigenvalues and eigenvectors, often the smallest ones
- Many eigenvalue problems in a sequence

Eigenvalue problem

A is real symmetric or Hermitian, λ is the eigenvalue, x is the eigenvector

$$Ax = \lambda x$$

Available tools

- LAPACK, EISPACK, PEIGS,...
- <u>Lanczos methods</u>
- Arnoldi method
- Davidson method
- minimizing Rayleigh quotient: <u>CG</u>,...

Lanczos algorithm

Building an orthogonal basis
$Q \equiv [q_1, q_2, \ldots, q_m]$

$$q_i = \frac{r_{i-1}}{\|r_{i-1}\|}$$

$$\alpha_i = q_i^T A q_i$$

$$r_i = A q_i - \alpha_i q_i - \beta_{i-1} q_{i-1}$$

$$\beta_i = \|r_i\|$$

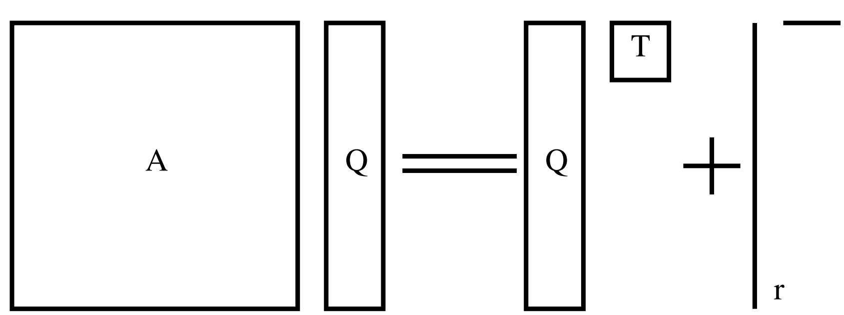
Rayleigh-Ritz projection

$$T = Q^T A Q$$

let $T = YDY^T$ be an eigenvalue decomposition of T

$$\lambda \sim d_{11},\ x \sim Q y_1$$

Lanczos algorithm



$$T = \begin{bmatrix} \alpha_1 & \beta_1 & 0 & 0 & 0 \\ \beta_1 & \alpha_2 & \beta_2 & 0 & 0 \\ 0 & \beta_2 & \alpha_3 & \beta_3 & 0 \\ 0 & 0 & \beta_3 & \alpha_4 & \beta_4 \\ 0 & 0 & 0 & \beta_4 & \alpha_5 \end{bmatrix}$$

Characteristics of the Lanczos method

- Advantages

  only need to access matrix through Aq

  few arithmetic operations per step

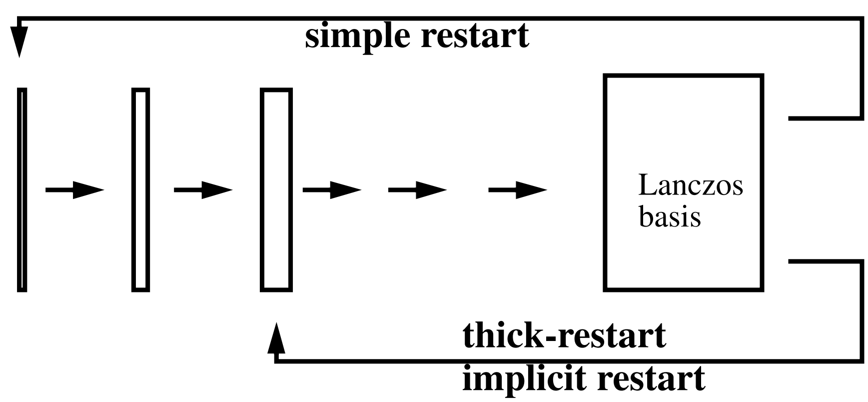
  effective for compute small number of extreme eigenvalues/eigenvectors
- Disadvantages

  need to use all Lanczos vectors -- unknown storage requirement

  degenerate eigenvalues do not converge at the same time

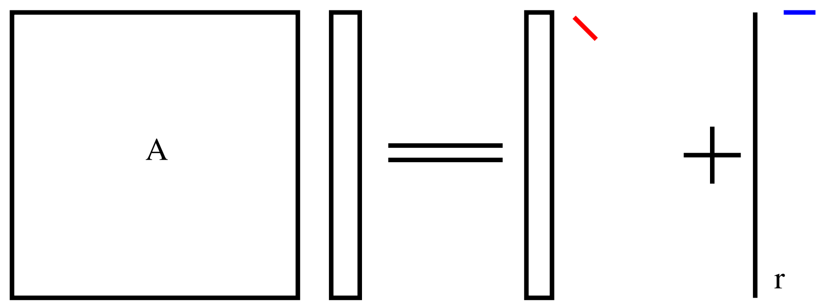
  only use one starting vector

Restarting the Lanczos algorithm

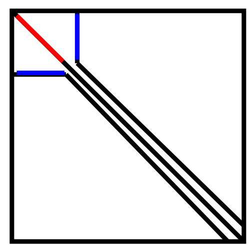


Thick-restart Lanczos

Restarting



New T after more steps



Thick-restart Lanczos

compared to non-restarted Lanczos method

- Use prescribed amount of memory(m Lanczos vectors)
- Effective restarting technique -- mathematically equivalent to implicit restarting
- same amount of arithmetic operations per step

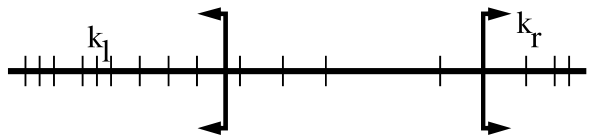
compared to implicitly restarted Lanczos method

- Easier to implement -- no bulge chase
- Compute Ritz pairs as in standard Lanczos method -- no extra postprocessing
- new dynamic restarting strategies

How to restart

Approximate deflation (Morgan, 1996)

Saving Ritz values near the wanted eigenvalue approximately deflates the spectrum, increases the separation and increases convergence rate



- A simple scheme:

  To compute a few smallest eigenvalues, user specify k_l ($k_r = m + 1$) and k_l smallest Ritz pairs are always saved when restarting
- starting point of dynamic schemes:

  test all possible choices of fixed k_l and observe the trend

Restarting heuristics

to achieve the performance of the optimal fixed thickness scheme without trying all possible choices

restart 1. empirical formula for k_l and k_r

restart 2. save those with small residual norms

restart 3. maximize the residual norm reduction in each step

restart 4. maximize the residual norm reduction of each restart loop

Test problems

Chelikowsky, et al., U of Minn

ab initio pseudopotential simulation of silicon clusters (*the first SCF step*)

- si4 4451x4451, 4-silicon cluster (12 smallest eigenvalues)
- si6 7949x7949, 6-silicon cluster (16 smallest eigenvalues)

Zunger, et al., NREL

empirical pseudopotential simulation of semiconductor materials

NOT self-consistent

- InGaP alloy, 512 atoms, 48x48x48 real space grid, 6603 planwave bases
- 9000-atom InGaAs quantum dot, 240 X 36 X 320 grid, 137,919 planwave bases

Comparison of restarting schemes

	si4			si6		
m	**20**	**50**	**100**	**20**	**50**	**100**
LANSO/locking	14.3	6.2	7.0	101.8	34.9	30.3
ARPACK	10.1	7.0	11.5	155.6	20.7	31.0
optimal fixed-k	5.2	3.2	4.6	50.0	7.9	11.9
restart 1	4.4	3.0	4.7	34.1	7.4	16.1
restart 2	4.9	3.1	4.6	28.4	7.4	16.0
restart 3	4.7	3.4	4.6	51.2	24.7	16.2
restart 4	5.0	4.0	6.8	49.4	17.9	19.3

Time (sec) on R10000

Comparison of restarting schemes

	si4 (243)			si6 (253)		
m	**20**	**50**	**100**	**20**	**50**	**100**
LANSO/locking	1729	715	758	4609	1761	1479
ARPACK	523	308	343	3373	421	471
optimal fixed k	488	274	268	1621	274	271
restart 1	504	296	297	1395	277	415
restart 2	543	296	297	1219	277	415
restart 3	384	286	297	1822	577	418
restart 4	439	286	294	1822	524	396

matrix-vector multiplications

Comparison against non-restarted Lanczos

512-atom InGaP alloy, 48 X 48 X 48 grid, 6603 G

Neig	TRLan	PLANSO
1	2.0	1.1
5	6.8	6.7
10	11.4	11.8
20	11.2	12.5
50	29.7	70.4
100	52.7	138.5

time (seconds) to compute the smallest eigenvalues on 8PE

Comparison against non-restarted Lanczos

9000-atom InGaAs quantum dot, 240 X 36 X 320 grid, 137,919 G

Neig	TRLan	PLANSO
1	72.0	59.0
10	164.5	142.4
20	184.8	172.9
100	612.0	--

time (seconds) to compute the smallest eigenvalues on 32 PE

Conclusions

- Effective method for computing large number of eigenvalues
- Efficient parallel algorithm, software available
- Good algorithmic scalability
- Fast restarting strategies (faster than ARPACK)