Shift Placement in a Parallel Spectrum Slicing Method for Symmetric Self-Consistent Eigenvalue Computation

David B. Williams-Young, Paul Beckman, Chao Yang

Supported by DOE SciDAC partnership project "Advancing Catalysis Modeling: From Atomistic Chemistry to Whole System Simulation" (PI: Martin Head-Gordon)

Introduction

Calculating a large number of the lowest eigenvalues and corresponding eigenvectors of a Hermitian matrix pencil is a common bottleneck in density functional theory codes. It is not easy to make traditional eigensolvers scale beyond thousands of processors. We present a spectrum slicing approach to compute these eigenpairs at each self-consistent field (SCF) iteration using subspace iteration on a number of shifted and inverted problems simultaneously. The key to achieving fast convergence and scalable performance is to partition the spectrum in an optimal fashion (without knowing where eigenvalues are in advance) and select shifts as centroids of eigenvalue clusters. We will show how this can be done throughout the SCF iterations [6].

The SCF Eigenvalue Problem

The SCF eigenvalue problem consists of the partial diagonalization of a convergent sequence of symmetric matrix pencils

$$F^{(i)}C^{(i)} = SC^{(i)}E^{(i)} \qquad i \geq 0$$

Where $F^{(i)}, S \in \mathbb{R}^{N\times N}$ are the Fock and overlap matrices corresponding to the *i*-th SCF iteration, respectively. $C^{(i)} \in \mathbb{R}^{N\times k}$ and $E^{(i)} \in \mathbb{R}^{k\times k}$ are the matrices of the *k* desired eigenvectors and eigenvalues of $(F^{(i)}, S)$, respectively. As the SCF iterations progress, $F^{(i)}$ converges towards a static matrix F.

Diagonalization of $(F^{(i)}, S)$ constitutes a computational scaling bottleneck in large scale electronic structure calculations. Typical methods adopted by the electronic structure calculations include:

Direct Diagonalization:

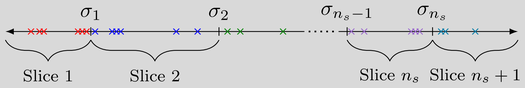
- Typically used when *k* is a considerable fraction of *N* ((e.g. Gaussian basis calculations)
- Scales to thousands of CPU cores for N = O(100,000).
- Scalability on a large number of computational resources (e.g. GPU accelerators) is limited.

Krylov-Subspace Methods:

- Typically used when *k* << *N* ((e.g. Plane wave calculations).
- Able to compute a small number of eigenpairs for N = O(1e6) or larger in only a few wall clock minutes.
- Ill-performant for *k* larger than a few thousand due to the need to solve a large projected eigenvalue problem directly.

Shift-Invert Spectrum Slicing

Spectrum slicing [1] for symmetric eigenvalue problems may be achieved by selecting a shift set, $\{\sigma_j\}_{j=1}^{n_s}$, which partitions the spectral region of interest into *spectral slices*.



In shift-invert spectrum slicing, we aim to compute eigenpairs in the spectral neighborhood of each shift. This may generally be forming a basis by the *shift-invert subspace iteration*

$$V_{(m+1)} = \operatorname{orth}\left(\left(F-\sigma_j S\right)^{-1} V_{(m)}, S\right), \qquad V_{(m)} \in \mathbb{R}^{N\times k}$$

Approximate eigenpairs may then be extracted via the Rayleigh-Ritz procedure. The basis along with its associated shift is known as a *spectral probe*.

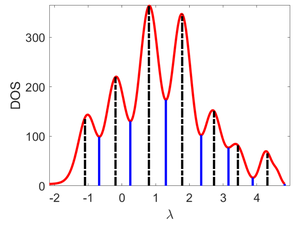
Spectrum Partition via Lanczos DOS

To determine initial shift placement, we estimate the density of states (DOS) of $(F^{(0)}, S)$ as a linear combination of Gaussian functions

$$y(\lambda) = \sum_{i=1}^{k} \zeta_i \exp(-\alpha_i(\lambda-\theta_i)^2)$$

Where ζ_i, α_i, and θ_i are related to the Ritz pairs extracted from the Lanczos procedure [2,5].

Initial shift placement is taken to be the local maximizers of the estimated DOS.

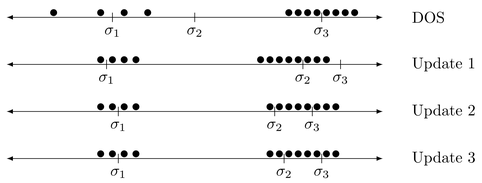


Shift selection via K-means Clustering

If the change in eigensystem between and is reasonably small, we may update the shift placement to more optimal locations via k-means clustering [3] of the validated eigenpairs of the previous SCF iteration.

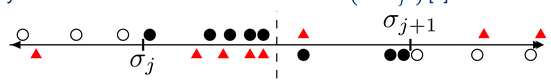
Clustering in our method serves two purposes:

- To determine more optimal shift placement after DOS shift selection
- To track the migration of eigenvalues throughout the SCF procedure



Spectral Slice Validation

Once approximate eigenpairs have been computed around each shift, each spectral slice may be validated by comparing the exact number of eigenpairs and number of approximate eigenpairs computed within each slice. The exact result may be computed through use of *Sylvester's Inertia theorem* on the factorized $(F-\sigma_j S)$ [4].



- In the case when adjacent spectral probes provide approximation for the same eigenpair (spectral overlap), the approximation from the probe which is closest to the eigenpair is selection as a validation candidate.
- If the number of validation candidates is the same as the exact number of eigenpairs within the spectral slice, the slice is *valid*.
- If the number of validation candidates is more than expected, those with the smallest residual are kept as valid eigenpair approximations
- If the number of validation candidates is less than expected, a probe must be added at the midpoint to obtain approximations for the missing eigenpairs.

Fixed Matrix

Spectral slicing of Silane ($SiOSi_5$) at the HF/6-31G(d) level of theory (*N*=1109). These examples exhibit our method in the regime of asymptotic convergence (fixed matrix). Each "SCF" cycle used 4 shift-invert subspace iterations on a basis of dimension 200. Each iteration is seeded with the basis of the previous iteration. These results compare the DOS and k-means shift placement schemes.

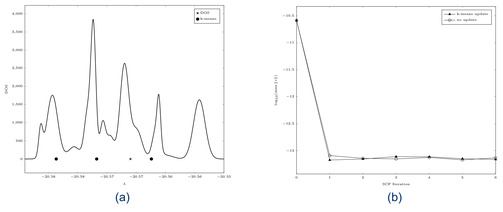


Figure 1: Spectral slicing of an eigenvalue cluster pertaining to core electrons ([-20.6,-20.55], 37 eigenvalues). (a) Shift placement according to DOS and k-means schemes, (b) convergence profile for the slowest converging eigenvalue in the energy window, (c) convergence profile using the DOS shift placement (no k-means update), (d) convergence profile using the k-means update. No noticeable difference between the two schemes for this energy window.

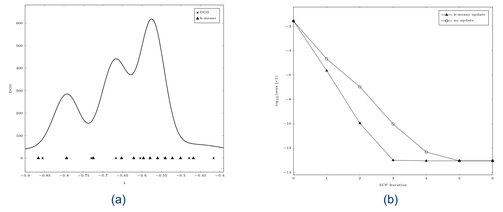
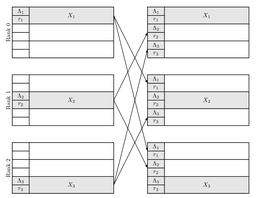


Figure 2: Spectral slicing of an eigenvalue cluster pertaining to electrons near the fermi level ([-0.9,-0.4]. 141 eigenvalues). (a) Shift placement according to DOS and k-means schemes, (b) convergence profile for the slowest converging eigenvalue in the energy window, (c) convergence profile using the DOS shift placement (no k-means update), (d) convergence profile using the k-means update. K-means shift placement is discernably beneficial for this energy window.

Parallel Implementation and Scaling

- Spectral probes may be constructed independently (and are thus are easily distributed)
- Slice validation scheme and eigenvalue clustering require synchronization between adjacent spectral probes



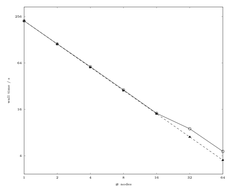


Figure 4: Strong scaling of our method for Silane HF/6-31G(d) (N=1109). Results were obtained On a Cray XC40 (Cori, Haswell compute nodes, 32 ranks / node). The open circles represent the overall wall time for diagonalization as a function of computational nodes, and the triangles represent the wall time required to form the spectral probes; the difference is the communication for probe synchronization. The dashed line represents linear scaling.

Dynamic Matrix

Spectral slicing of Silane ($SiOSi_5$) at the HF/6-31G(d) level of theory (*N*=1109). This example exhibits our method applied to a convergent sequence of matrix pencils (14 iterations). Each SCF cycle used 4 shift-invert subspace iterations on a basis of dimension 200. Each iteration is seeded with the basis of the previous iteration. These results compare the DOS and k-means shift placement schemes.

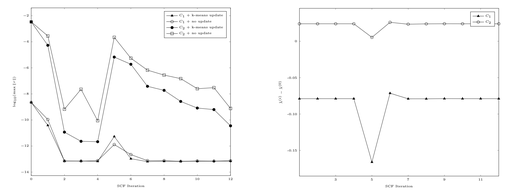


Figure 3: Convergence behavior for the spectral slicing of Silane for a convergent SCF sequence. (a) shows the comparison of DOS and k-means shift selection schemes throughout the SCF procedure. C_1 and C_2 represent the same energy windows as Figures 1 and 2, respectively. (b) Shows the change in average (exact) eigenvalues for these clusters throughout the SCF procedure. There is a considerable change in the eigenspectrum in iteration 5 of the SCF procedure as seen both in (a) and (b). K-means exhibits superior convergence behavior for C_2 and roughly the same behavior for C_1, as would have been expected according to Figures 1 and 2.

References

[1] R.G. Grimes, J.G. Lewis, and H.D. Simon, SIAM Journal on Matrix Analysis and Applications, vol. 15, no. 1, pp. 228–272, 1994.

[2] R. Li, Y. Xi, E. Vecharynski, C. Yang, and Y. Saad, SIAM Journal on Scientific Computing, vol. 38, no. 4, pp. A2512–A2534, 2016.

[3] S.P. Lloyd, IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129–137, 1957.

[4] J.J. Sylvester, Philosophical Magazine, vol. 4, no. 23, pp. 138--142.

[5] Y. Tang, D. B. Williams-Young, Uncertainty quantification for density of states estimation using Gaussian processes, In preparation

[6] D. B. Williams-Young, Shift selection in a parallel spectrum slicing algorithm for self-consistent eigenvalue computation, In preparation