New eigensolvers for large-scale nanoscience simulations

A Canning\(^1\), O Marques\(^1\), C Voemel\(^1\), Lin-Wang Wang\(^1\), J Dongarra\(^2\), J Langou\(^2\) and S Tomov\(^2\)
\(^1\)Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
\(^2\)Dept. of Computer Science, University of Tennessee, Knoxville, TN 37996, USA

E-mail: acanning@lbl.gov

Abstract. We present results for applications to nanosystems of state-of-the-art iterative eigensolvers based on conjugate gradients and variants of Davidson in the context of semi-empirical plane wave electronic structure calculations. We are concerned with the computation of electronic and optical properties of nanosystems using the Energy SCAN method to compute interior eigenstates around the band gap that determine their properties. Numerically, this interior Hermitian eigenvalue problem poses several challenges, with respect to both accuracy and efficiency. All the iterative eigensolvers are seeking the minimal eigenvalues of the folded operator with reference shift in the band-gap. The tested methods include standard conjugate-gradient (CG)-based Rayleigh quotient minimization, Locally optimal block-preconditioned CG (LOBPCG) and two variants of the (Jacobi-)Davidson method: JDQMR and GD+1. Our experimental results indicated that Davidson-type methods are often the fastest.

1. Introduction

The computation of the electronic properties of large nanostructures such as quantum dots is an important field of current research. Its significance is highlighted by a number of activities such as the DOE-funded SciDAC project Predicting the Electronic Properties of 3D Million-atom Semiconductor Nanostructure Architectures. In this paper we focus on the mathematical aspects of this project. Some of the physics applications of this project are discussed in another publication in this volume [1].

We are interested in electronic structure calculations based on the solution of an effective single-particle Schrödinger equation,

\[ H\psi_i(r) = \left(-\frac{1}{2}\nabla^2 + V(r)\right)\psi_i(r) = E_i\psi_i(r), \]

where \(H\) denotes the approximate Hamiltonian. The atomic system is described by the potential \(V(r)\), which we assume as an externally given empirical pseudopotential [2]. The set \(\psi_i(r)\) denotes the orthogonal wave-functions (eigenstates) and \(E_i\) their corresponding energies. In the context of the self-consistent field iteration [3], a large number of eigenstates corresponding to all the electron states of the system need to be computed. On the other hand, only a small number of eigenstates are relevant for determining certain optical and electronic properties. Our task is to compute these states at the top of the valence and bottom of the conduction band. These states typically lie in the middle of the spectrum with too large a distance from the lowest (ground) state to efficiently compute all of them from the lowest up.

In a discrete, finite plane-wave basis, the single-particle Schrödinger equation directly translates into a Hermitian eigenvalue problem, and we use the same notation \(H\) for the matrix in question. The discrete Laplacian is the commonly computed in the plane-wave basis because it is diagonal. On the
other hand, the discretized potential is diagonal in real space. Thus the resulting \( H \) becomes implicitly available only through matrix-vector products and via the fast Fourier transform (FFT). This together with the large system size requires the use of efficient iterative eigenvalue methods.

Our parallel ESCAN code [4] uses a folded spectrum approach [2] to find the interior eigenstates that we are looking for. Based on physical knowledge about the system, a reference energy \( E_{\text{ref}} \) was chosen, and then the smallest eigenvalues of the system \( (H - E_{\text{ref}})^2 \) were computed via the PCG (parallel conjugate gradient) Rayleigh quotient. In our study, we compare this method with blocked CG solvers and variants of the (Jacobi-)Davidson method.

2. Overview of eigenvalue methods

The eigenvalues of \( H \) near \( E_{\text{ref}} \) can be ideally tackled by a shift-and-invert strategy, which conveniently maps those eigenvalues into the dominant and well-separated eigenvalues of \( (H - E_{\text{ref}} I)^{-1} \), while the eigenvectors remain unchanged. (In practice, operations involving \( (H - E_{\text{ref}} I)^{-1} \) are replaced by solutions of systems of linear equations through direct methods.) However, such a strategy is infeasible in the present context because the size of our systems makes the number of operations required for shift-and-invert too costly. In addition, in our codes the matrix \( H \) is never explicitly available: we can only afford to obtain the action of \( H \) on a vector. One may also consider the use of polynomial spectral transformations on \( (H - E_{\text{ref}} I) \) to enhance spectral properties of interest. The folded spectrum method (FSM) [2] is a typical illustration of this approach. However, it remains to be verified whether polynomials of low degree are the best approach to capture information about interior eigenstates.

A nonlinear CG algorithm can be used to successively minimize the Rayleigh quotient function \( f(x_i) = \langle x_i, Hx_i \rangle / \langle x_i, x_i \rangle \), where the gradient is given by \( \nabla f(x_i) = Hx_i - x_i (x_i^* Hx_i) (x_i^* x_i) \). (The minimum of the Rayleigh quotient corresponds to the minimum eigenvalue of \( H \).) The method can also be implemented in a blocked form, whereby the minimization is simultaneously applied to a set of orthogonal vectors \( X_i \). Together with a suitable preconditioner and FSM, this approach (PCG) has been used to solve a number of practical problems of interest. A potential improvement over the PCG algorithm is the locally optimal preconditioned conjugate gradient (LOPCG) method [5], which extends PCG by applying Rayleigh Ritz on \( \text{span}(w_i, x_i, x_{i-1}) \), where \( w_i = P \nabla f(x_i) \) and \( P \) is the preconditioner. As before, the method is amenable to an implementation by blocks (LOBPCG).

A class of algorithms based on subspace or projection-based techniques can also be used to seek eigenstates of interest. This class includes the algorithms of Lanczos and Arnoldi. Starting with a vector \( q \), these algorithms generate an appropriate basis for a Krylov subspace. Often the trade-off of these methods lies between the sizes of the basis. While the dimension of the subspace is usually much smaller than the dimension of the original problem, if we can afford to set it larger, then better approximations for the eigenvalue and eigenvectors are obtained as the computation progresses. However, finite-precision arithmetic requires explicit orthogonalization of the vectors generated, which implies significant cost for large subspaces. We note the Krylov-based techniques can also be implemented in blocked form. Often, subspaces of small dimensions do not produce good approximations for the, say \( k \), eigenstates of interest. Then, some kind of restarting strategy with a proper set of vectors has to be employed, possibly a number of times until convergence for the wanted solutions is obtained. The ARPACK implementation of the Arnoldi algorithm, for example, first generates a subspace \( S \) of dimension \( k+p \). Then, information obtained from the associated reduced problem \( T \) (i.e., the projection of \( H \) into \( S \), which also has dimension \( k+p \)) is used to project \( S \) into another subspace of dimension \( k \). In order to do this, the eigenvalue approximations obtained from \( T \) that are not relevant, \( \theta_n \), are used as shifts for a QR factorization of \( T \), i.e. \( T - \theta_n I = QR \), with \( Q \) unitary and \( R \) upper triangular. The resulting leading \( k \) columns of \( Q \) are used for reducing the dimensions of \( T \) and \( S \), from which point the algorithm is restarted. The goal is to make the (restarting) subspace of dimension \( k \) poor in the directions of the irrelevant solutions, therefore promoting convergence to the sought solutions. This type of restart is referred to as implicit in the sense that spectral information is embedded into the restarting subspace. It has already enabled the computation of eigenstates of large
A variant of this restarting scheme is implemented in the thick-restart Lanczos method, which employs a subspace of dimension \( k \) for restart (as above), but which is explicitly built with Ritz vectors (eigenvectors approximations) obtained from \( Q \).

More recently, variants of Davidson’s method have been shown to be effective for solving a number of applications. Davidson’s original algorithm targeted mainly the smallest eigenvalues of (strongly diagonal dominant) matrices arising from applications in Chemistry. It relies on the fact that in the case of a strongly diagonally dominant matrix, the inverse of its diagonal entries is a good approximation for the inverse of the matrix (and thus a good preconditioner). In this case, given an approximate eigenpair \((x, \theta)\) and the residual \( r=(\theta I-A)x \), the vector obtained from \( t=(\theta I-\text{diag}(A))^{-1}r \) is used to generate a new vector to be added to the subspace \( S \). Depending on \( A \) and \( \theta \), however, the preconditioner \((\theta I-\text{diag}(A))^{-1}\) may be ill conditioned. To remedy this situation, an orthogonal component correction has been proposed, which solves \((I-xx^*) (A-\theta I)(I-xx^*)t=-r\) approximately (by means of the QMR algorithm for example). This is the essence of the Jacobi-Davidson algorithm. It allows for the generation of more general subspaces than the original algorithm and takes great advantage of the preconditioner.

3. Test results
We perform experiments on a 1,061-atom cadmium-selenium quantum dot and a 66,624 atom InAs quantum wire. The size of the Hamiltonian matrix for the quantum dot is 141,625 and 2.7 million for the quantum wire. The goal of the experiments is to compare the different iterative methods in terms of time to solution and robustness. The methods compared are a set of two conjugate gradient-like methods Banded-PCG and LOBPCG, PARPACK, and two (Jacobi-)Davidson-like methods GD+k (Olsen’s algorithm) and JDQMR as implemented in the PRIMME [6] package. All of these methods are extremely robust. All the runs are performed on 32 processors of an IBM SP5 at the NERSC computing center. Various sets of parameters were tested for GD and JDQMR in all the cases, and we present the results for the best parameters.

Table 1: Matrix vector multiplications and time for calculation of 10 conduction band minimum (CBM) states for a 1061-atom CdSe quantum dot using a nonlocal empirical pseudopotential. The * indicates that the algorithm converged but missed relevant eigenvalues.
Table 2: Matrix vector multiplications and time for calculation of 5 valence band maximum (VBM) states for a 66,624-atom InAs quantum wire using a local empirical pseudopotential. The * indicates no convergence with the maximum number of allowed restarts.

| Algorithm (Folded) | Matrix-Vector Multiplications | Time (s) |
|--------------------|-------------------------------|----------|
| Band-by-Band CG    | 149726                        | 7278     |
| LOBPCG             | 56207                         | 3690     |
| PARPACK            | *                             | *        |
| GD+k               | 26326                         | 1424     |
| JDQMR              | 81364                         | 3877     |
| Unfolded           |                               |          |
| PARPACK            | *                             | *        |
| GD+k               | 39828                         | 2452     |
| JDQMR              | 56441                         | 2746     |

As can be seen from tables 1 and 2, GD+k gives the best results for the folded spectrum method and for the unfolded spectrum for the larger quantum wire. It can be five times faster than previously used methods. Our experiments have also indicated that although PARPACK may exhibit fast convergence in the unfolded spectrum for some cases (see table 1), it may miss eigenvalues of interest. A more detailed discussion of some of the results in this paper, along with more results on CdSe quantum dots of different sizes, can be found in [7].

As part of this project we have also developed preconditioners based on using a modified form of the solution of the bulk system of the nanostructure we wish to study. The theoretical basis for the success of this method is that, in the interior of the nanostructure, the properties of the system are bulk-like so the bulk solutions are a good approximation to the true solution in this region. The bulk solutions are easily derived in a negligible amount of time. We have implemented the bulk-band preconditioner in the PCG version of ESCAN (see [8] for details) and are in the process of implementing this preconditioner for the other solvers.

The parallel ESCAN code is available for download from http://icl.cs.utk.edu/doe-nano/software/pescan. This website gives more information about the results discussed here, and also a set of benchmark systems.

Acknowledgments
This work was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing (MICS) and Basic Energy Science under LAB03-17 initiative (Theory, Modeling and Simulation in Nanoscience), contract Nos. DE-FG02-03ER25584 and DE-AC03-76SF00098. We also used computational resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy. We thank A. Knyazev and A. Stathopoulos for their advice regarding the use of the LOBPCG and the PRIMME packages.

References
[1] Wang L W 2008 Large scale ab initio nanostructure electronic structure calculations for energy applications Proc. SciDAC 2008
[2] Wang L W and Zunger Q 1994 Solving Schrödinger’s equation around a desired energy: Application to silicon quantum dots J. Chem. Phys. 100: 2394
[3] Payne MC, Teter M P, Allan D C, Arias T A and Joannopoulos J D 1992 Iterative minimization techniques for ab initio total-energy calculations: Molecular dynamics and conjugate
gradients Rev. Mod. Phys. 64(4):1045

[4] Canning A, Wang L W, Williamson A, and Zunger A 2000 Parallel empirical pseudopotential electronic structure calculations for million atom systems. J. Comp. Phys. 160:29

[5] Knyazev A V 2001 Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method SIAM J. Sci. Comput. 23(2):517

[6] Combs J R and Stathopoulos A 2006 PRIMME: Preconditioned iterative multiMethod eigensolver: Methods and software description (Technical Report WM-CS-2006-08 Computer Science Dept. The College of William and Mary)

[7] Voemel C, Tomov S, Marques O, Canning A, Wang L-W and Dongarra J 2008 State-of-the-art eigensolvers for electronic structure calculations of large scale nano-systems J. Comp. Phys. (to appear)

[8] Voemel C, Tomov S, Wang L-W, Marques O and Dongarra J 2007 The use of bulk states to accelerate the band edge state calculation of a semiconductor quantum dot J. Comp. Phys. 223:774