The quantitative applicability of subspace optimization with localization to N deep level defects in SiC

David Raczkowski
NERSC, Lawrence Berkeley National Laboratory, Berkeley, CA 94720.

C.Y. Fong
Department of Physics, University of California, Davis, California 95616-8677

We investigate the applicability of a subspace optimization method, which results in localized nonorthogonal orbitals that are restricted in space, modified for a singly occupied deep level defect. We use the substitution of a silicon atom by a nitrogen atom in SiC for our test purposes. The calculations use the serial Gaussian DFT code SEQQUEST within the LDA approximation. The approximation of the localization of the orbitals gives linear scaling of the dominant parts of the algorithm thus allowing for large systems to be performed on a workstation. Our intent is to determine the ease of such a method with which quantitatively accurate calculations may be done for a deep level defect system.

PACS numbers:

I. I. INTRODUCTION

Linear scaling[1],[2] algorithms and occupied subspace optimizations[3],[4] incorporating localization have typically allowed only for full occupancy of the orbitals. All of the localized orbitals are usually treated equally with no differentiation as to the occupancy of the orbitals. We present an algorithm that incorporates partial occupancy and investigate the applicability to quantitative calculations involving deep level defects. A deep level defect is a good starting point for generalizing these algorithms to systems with partial occupancies, such as metallic states, since only the defect-state is partially occupied and extended. The defect level eigenfunction is relatively extended when compared to the localized orbitals, which result in defect-free areas of a semiconductor or insulator crystal.

Defects of materials such as silicon carbide are important for understanding real world properties of non-pure materials. They are also important in understanding the properties of doped materials. In both cases, a large number of atoms is necessary to investigate defect-defect interactions and many different doping concentrations. Therefore it is desirable to extend recent algorithmic developments[3], previously restricted to systems with a gap. We use a N substitution of Si in SiC as a system to test the algorithm.

We begin our discussion with a brief of review of past work along with the simple modifications for incorporating the partial occupancy of the defect levels. We then proceed to discuss the properties of the defect system obtained from calculations with no localization (actually using diagonalization). Next, we examine the effects of different localizations on the accuracy of the final positions of the atoms and the energetics for these configurations. Here, we discover that we achieve high accuracy by allowing the orbitals near the defect to be of much longer range than those away from the defect, and thus providing accuracy where needed. We end with a discussion on local minima in the minimization of the total energy and the impact on the way one should proceed in such calculations.

II OCCUPIED SUBSPACE OPTIMIZATION

Only a general overview of the minimization method is given so that we may introduce the modifications for partial occupancy. For a more detailed discussion see Ref [3]. For each self-consistent iteration, instead of solving iteratively or exactly the generalized eigenvalue equation,

$$\mathbf{H\Psi{=}S\Psi E}, \tag{1}$$

the trace of a generalized Rayleigh quotient is minimized

$$Tr[(\mathbf{\Phi^\dagger S\Phi})^{-\mathbf{1}}\mathbf{\Phi^\dagger H\Phi}] \tag{2}$$

The orbital matrix, $\mathbf{\Phi}$, is an M x N matrix of the coefficients of the M basis functions for the N localized nonorthogonal orbitals. We implement a Grassmann conjugate gradient (GCG) algorithm[5] to minimize this trace. For systems which have all eigenvectors fully occupied the density matrix is simply calculated by using

$$\mathbf{P} = \mathbf{\Phi L \Phi^\dagger}, \tag{3}$$

where

$$\mathbf{L} = (\mathbf{\Phi^\dagger S\Phi})^{-\mathbf{1}}. \tag{4}$$

$\mathbf{P}$ is then used to calculate the charge density. Note that $\mathbf{L}$ is the density matrix within the occupied space defined by $\mathbf{\Phi}$, while $\mathbf{P}$ is the density matrix with respect to the M basis functions.

For partially occupied systems, the eigenvectors near the gap have to be found. Let us now assume that the N lowest eigenvectors are fully (i.e. doubly due to spin) occupied and the N+1 eigenvector is singly occupied. $\mathbf{L}$ is now split into two separate contributions. The first $\mathbf{L}_1$

from the fully occupied orbitals and the second $\mathbf{L}_2$ from the singly occupied defect-state eigenfunction.

$$\mathbf{L} = \mathbf{L}_1 + 0.5\mathbf{L}_2 \tag{5}$$

This can be done in several different ways. We accomplish this in the same manner as Ref.[6] by solving the smaller generalized eigenvalue equation in the space of K localized non-orthogonal orbitals where $K \geq N{+}1$,

$$(\mathbf{\Phi}^\dagger \mathbf{H} \mathbf{\Phi})\mathbf{\Psi_K} = (\mathbf{\Phi}^\dagger \mathbf{S} \mathbf{\Phi})\mathbf{\Psi_K}\mathbf{E_K}. \tag{6}$$

In this work, we use K equal to $N{+}4$. The result is

$$\mathbf{L}_1 = \sum_{i=1}^{N} \Psi_i \Psi_i^* \tag{7}$$

and

$$\mathbf{L}_2 = \Psi_{N+1}\Psi_{N+1}^*. \tag{8}$$

This algorithm adds an additional $O(N^3)$ step of diagonalizing Eq. (6) to the two previous $O(N^3)$ steps of calculating $(\Phi^\dagger \mathbf{S}\Phi)^{-1}$ and the multiplication of $(\Phi^\dagger \mathbf{S}\Phi)^{-1}$ $(\Phi^\dagger \mathbf{H}\Phi)$ when all orbitals are equally occupied. The $O(N^3)$ scaling is not a concern for system sizes up to 1,000 atoms.[3] Once the system size demands a truly linear scaling algorithm, one must modify the above method for calculating singly occupied states. In this case, an alternative[7] is to solve for the space of K localized orbitals, and then find the largest eigenvectors of this space. In cases where the defect level is very close to the unoccupied bands, it may be more efficient to solve for the N lowest states as a block and solve for the $N{+}1$ state with a separate minimization algorithm while enforcing orthogonalization to the lowest N orbitals. These algorithms would scale linearly as a fixed number of eigenvectors are found. In either situation, the defect-state eigenvectors are weighted properly and their contribution to $\mathbf{L}_2$ (each calculated via Eq. 7) are respectively subtracted from or added to $\mathbf{L}_1$, which is calculated via Eq. 4.

In order to achieve linear scaling, each localized nonorthogonal orbital, a column of Φ, is restricted to have a non-zero contribution from only selected Gaussian basis functions. One inputs the localization radius for the single zeta, double zeta, and polarization shells, e.g. {5;5;5}. This is an approximation that is exact in the limit that all of the basis functions are used. The computational effort of the dominant parts scale as O(N), with the use of sparse matrix multiplies, and at some system size (crossover point) is more efficient than diagonalization. One takes the advantage that the desired accuracy may be achieved by concentrating the computational effort where the interactions are the strongest, and hopefully with a relatively small (compared to M) number of basis functions.

III. DEFECT CALCULATIONS USING DIAGONALIZATION

We use a N substitution for a Si within cubic SiC as our test system for the algorithm to handle the partially occupied defect levels. We chose SiC, a wide gap semiconductor that can be operated at high temperature and high pressure, for its technological importance[8] as well as prior knowledge of its localization properties.[3] If the system is started in the T_d symmetry (ideal cubic positions) then the final singly occupied state is essentially triply degenerate. If the system is perturbed into a C_{3v} symmetry - N shifted away from a nearest neighbor (n.n.) atom - then the degeneracy is broken. We use a double-zeta with polarization basis set along with norm-conserving pseudopotentials and the gamma point for our **k**-point sampling.

For a 64-atom SiC system in the cubic phase, we look at two configurations of a nitrogen atom replacing one Si atom. Our purpose is to compare the accuracy of the forces and energies for this system. Since we cannot compare the energies of the diagonalization directly with the calculations using localization, we compare the relative energies of these two defect systems. In configuration (C1) as depicted in Fig. 1, nitrogen directly takes the position of a Si atom at the origin. In (C2) as depicted in Fig. 2, the nitrogen atom is placed as in (C1) and then switched with the furthest n.n. C in the C_{3v} symmetry, the C in the [111] direction in our systems. We perturb both configurations into the C_{3v}symmetry and relax the atomic positions. Figures 1 and 2 show the movement from the initial positions to the final positions for atoms around the defect for C1 and C2 respectively.

In (C1), the nitrogen moves slightly in the [-1,-1,-1] direction, the n.n. C at [1,1,1] moves significantly in the [-1,-1,-1] direction, and the other n.n. C move even closer to the N. The C2 configuration undergoes a more radical change in positions than C1. The C at the origin moves much closer to the other C while the N moves much closer to its neighboring Si. In (C2), the C (now at the origin due to the swap) moves significantly in the [-1,-1,-1] direction, oppositely the N moves in the [1,1,1] direction, and the other n.n. C move closer to the C that started at the origin. We use the final positions for these 5 atoms near the defect as the starting positions for the calculation using localization. Using these two configurations, we are able to investigate the issues of localization concerning two systems that are dissimilar locally (near the defect), but similar for the rest of the composition.

For the 64-atom unit cell, Table 1 gives the difference in the energy of the two configurations at the final relaxed atomic positions using diagonalization. C2 is lower in energy than C1. The difference in the eigenvalues of the lower-energy defect state to the now doubly-degenerate, higher-energy defect states (defect level gap) is more pronounced for C2 at 37mRy than C1 at 5.4 mRy. This is consistent with the greater shift in the atomic positions for C2, thus causing a greater splitting of the degenerate defect levels. For the C1 configuration, the lattice vectors remained cubic and were optimized to 23.052 Bohr from the 23.165 Bohr of the ideal crystal. The stress on the unit cell for the second configuration was small enough that no unit cell relaxation was warranted. As we are mostly interested in comparing the diagonalization and the optimization method, just using identical geometries is most important.

We next look at a 216-atom system for the two defect configurations described above. We use the unit cell of the ideal cubic SiC crystal for both configurations. The energy difference between the two configurations decreases as the system size increases presumably from a change in the defect-defect interactions from the defect in one unit cell to another. The C1 configuration undergoes a larger decrease in the defect-defect energy. This observation also suggests that the defect eigenstates for the C2 configuration are more localized. The higher localization and the larger defect level gap promote the concept of more ionic and stronger bonds for the C2 system, which might explain the relatively lower energy for C2. These observations are additionally consistent with the final atomic positions of C2 compared to C1. For the 216-atom system, the defect level gap decreases for C1 to 1.6 mRy and for C2 to 17 mRy. The results in Table 1 will be used as a reference to compare the accuracy of calculations with different localizations.

IV. DEFECT CALCULATIONS WITH LOCALIZATION

For the sparse calculations on the 64-atom unit cell, we place 4 orbitals on the nitrogen atom and 4 on the carbon atoms creating a space of 132 orbitals (128 of which are fully occupied) in which we diagonalize. We investigated having one orbital on the nitrogen but found this to be inferior. The starting atomic positions were the same as for the calculations using diagonalization. 8-12 geometry steps, essentially the same as for diagonalization, were typically sufficient to obtain a magnitude of any force smaller than $2\text{x}10^{-3}$ Ry/Bohr. We use a maximum of 15 GCG iterations with a convergence criteria of 10^{-10} divided by the number of orbitals for the square norm of the gradient. A smaller maximum number may be used, but smaller values typically cause slower self-consistent convergence and thus take a longer total time. Table 2 gives the relative energies for localization regions and settings that gave relaxed atomic positions.

For (C1), we first tried bond-centered orbitals including all basis functions from the closest 8 atoms. This setting did not relax the atomic positions to the desired accuracy. Since for larger localization regions atom-centered orbitals are more efficient[3], we switched to atom-centered orbitals and use them for all reported results. For the localization (S1) including all basis functions within 7 Bohr, we were able to relax the structure, but the error in the energy and the final atomic positions seemed insufficient. A localization (S2) with the nitrogen orbitals occupying the full variational space gave essentially the same relaxed atomic positions as with diagonalization for (C1). However, this setting did not give accurate relaxed atomic positions for (C2). This inaccuracy coupled with the very poor value for the energy difference suggested that (C2) needed more variational flexibility near the defect.

A localization (S3), with fully extended orbitals additionally on the n.n., gave accurate final atomic positions and relative energy for both configurations. Since the relative number of orbitals that are fully extended is rather small, the extra cost is not significant. We did investigate bond-centered orbitals with extended orbitals also near the defect. We found the atom-centered orbitals still to be more efficient and accurate. We also tried a localization (S4) in which the orbitals on second n.n. C were extended. The increased localization for the orbitals on the second n.n. does not seem to be beneficial. We found other instances where larger localizations regions also perform worse compared to select smaller one. A localization (S5) of 7 Bohr for the N and 1 n.n. C and 5.5 Bohr on the rest of the orbitals gave an accuracy much closer than that of the larger localization of S1. The number of geometry steps was larger for S5 though giving an almost equal time to S3.

Table 3 gives the relative energies between relaxed 216-atom unit cells of the two defect configurations. In order to decrease the computational effort, we started from the positions from the diagonalization calculations. The localization region of S3 gives a value within our desired accuracy of 1mRy. A value of $5\text{x}10^{-4}$ for *cut_gro*[3] (signifies accuracy of $\mathbf{S}\Phi$ multiplication) was used for all calculations. A calculation using the localization region of S1 for C1 with a growth parameter of 10^{-3}gave forces that did not monotonically decrease. The energy decreased, but as the forces jumped around this was seen as an inferior setting to $5\text{x}10^{-4}$, which gave monotonically decreasing forces.

For the 216-atom unit cell, we also looked at scaling back the cutoffs in order to increase efficiency. We start from the same positions as for diagonalization for this calculation. We use: (1) a grid spacing of 0.4095 (previously 0.273) Bohr for the solution of Poisson's equation and 10^{-6} (previously 10^{-8}) for *convgr* (cutoff value for a basis function to have a non-zero value on a grid point), (2) a factor of 100 larger cutoff criteria for the setup of $\mathbf{S}$ and $\mathbf{H}$, and (3) the localization region of S5. These settings are denoted as (S6). The benefit of increasing the setup cutoffs is mainly in the set up time, about 2 times faster. The matrices are slightly sparser, but not significantly so to shorten the time for the sparse multiplication. The time spent on the grid is about 4.25 times faster. The overall saving was about a speedup 4 times faster per geometry step. In terms of the accuracy, the difference is mostly attributable to the smaller localization region. With these settings and localization region, the subspace optimization method is faster by about 20% than diagonalization with comparable settings.

V. ISSUES FOR OBTAINING QUANTITATIVE ACCURACY I

Other papers report finding local minima[9],[10] for the minimization of Eq. (2). The local minima result from localization and are not found when no localization is used. In our results so far, we have not seen any egregious instances of local minima causing poor results. With very reasonable localization regions, we have been able to achieve very high accuracy. However, that is not to say that local minima have not been encountered, they have just been overcome in the process of relaxing the atomic positions. Towards the end of convergence, we do see the GCG algorithm stalling during one or more SCF steps. The proper search direction cannot be found due to localization. The result is that we also commonly find a stalling of the self-consistent minimization procedure. We now present an instance where problems occur due to "local minima".

In Table 4, we compare the energy obtained at the end of a relaxation to the energy obtained with a single total energy calculation at the final atomic positions. The energies are lower for the relaxation as the minimization procedure has started at later geometries with a Φ closer to the electronic energy minimum. A path for Φ is followed with the atomic movements allowing the GCG algorithm to break free from what could be called a local minimum. If the minimum is found too quickly, e.g. a single calculation at the final geometry, the GCG minimization is not allowed the time to slowly find the minimum or the ability to evade the " local minima". Depending on how far two configurations start from the final positions, the final relative energies may be inaccurate. We have found differences of a similar magnitude for other systems, specifically yttria-stabilized Zirconia. This may manifest itself in the low formation energies of YSZ, which comprised total energies obtained through relaxation of the atoms and total energies at a fixed geometry, in a recent paper.[11]

If the SCF procedure is restarted with the old guess for Φ, the energy is lowered. If done judiciously, one could retrieve the accuracy lost when comparing energies obtained through atomic relaxation and energies at a fixed geometry. In Table 5, we present the total energy obtained from restarting with the Φ from previous restarts. For example, we take the Φ obtained after one restart and use it as the starting point for the next calculation. The results show clearly the extent to which local minima are a problem for this calculation. If one desires accuracy of the order of 10 mRy then the local minima are not a problem. Even after 1 or 2 times restarting, we have recovered the lost energy and have a value comparable to that achieved with relaxation. This is not an ideal solution, as it is not known *a priori* how many restarts should be done.

A. Consistency

LCAO calculations inherently require more consistency in the calculation as the basis sets are not uniform as in plane wave calculations. Since we are looking at relative convergence the two calculations must be very similar for a cancellation of errors. This is also true for plane wave calculations with regards to using equivalent **k**-points. The **k**-point sampling is not necessarily fully converged, but the **k**-space integration is equally represented in the two calculations. Now with localization, one has to pay even more attention to consistency than in the typical LCAO calculation. The localization regions must be consistent. This is not straightforward for systems that have different atomic geometries. Ideally one should keep the same number basis functions for each type of orbital. This may require different localization radii. If there is a drastic difference in atomic positions this may not be the best course of action. One may just want to keep the localization radii the same.

One also now has to worry about the path to the electronic ground state. One has to ensure that the initial geometries are comparable so that a similar number of geometry steps are made. This will allow equal time for the GCG algorithm to overcome any critical slowdowns or local minima. One could also implement a self-consistent procedure by restarting with the final Φ once the final geometry is found. The reuse of orbitals obtained in a chemically similar or equivalent environments would overcome the problem in the same way the reuse of Φ helps from previous calculations. A solution,[12] which may overcome the need for such consistency in starting positions and the values of Φ, calculates a larger *occupied* subspace and then resolves the occupations within this space. This solution has the burden of requiring a significantly greater number of orbitals. More importantly it requires absolute convergence at every SCF step in order to preserve electron number. The best solution for overcoming "local minima" might be the transferability of orbitals from chemically similar environments.[13] In this fashion, the local minima might be averted without any additional cost.

V. SUMMARY

We have presented a method for performing DFT calculations using a Gaussian basis for systems that involve partially occupied defect levels. The computationally dominant parts of the algorithm scale linearly thus allowing for calculations of very large systems. We found that high accuracy could be achieved by allowing the orbitals near the defect to be long ranged while the other localized orbitals in the regular part of the solid to be relatively restricted in space. The efficiency crossover point was achieved for 216 atoms for a localization that still gave respectable accuracy (within 2%). We also have addressed some issues for obtaining quantitative accuracy with localization. The main concept is that consistency must be maintained at all levels (a much more difficult task than in standard calculations).

Acknowledgements

Partial support was provided by the Campus Laboratory collaboration grant of the University of California and Sandia National Laboratories. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockhead Martin Company, for the United States Department of Energy, under contract No. DEAC0494AL85000. C.Y. Fong acknowledges support from a NSF grant (no. Int-9872053). DBR acknowledges support and hospitality from Ford Motor Co. during the summer 1999. We gratefully acknowledge P. A. Schultz for stimulating conversations and his proposal of a defect system.

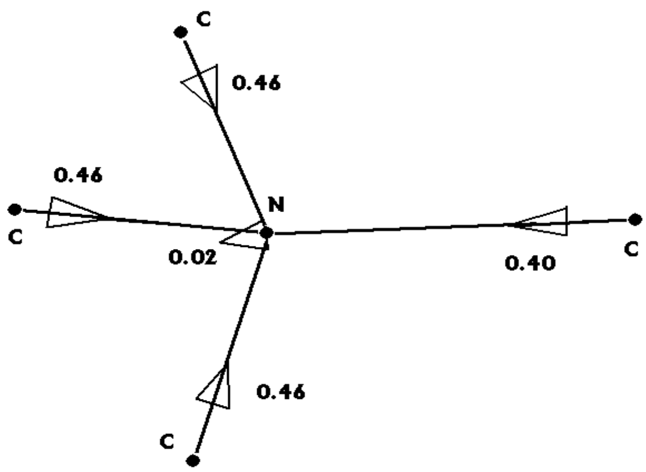

FIG. 1: Fig. 1 Movements from ideal cubic positions around N defect for configuration 1 (C1)

[1] F. Mauri and G. Galli, Phys. Rev. B **50**, 4316 (1994).
[2] P. Ordejón, D. A. Drabold, R. M. Martin, et al., Phys. Rev. B **51**, 1456 (1995).
[3] D. Raczkowski, C.Y. Fong, P.A. Schultz, R.A. Lippert, and E.B. Stechel, Phys. Rev. B **64**, 1555203 (2001)
[4] J. Bernholc, E.L. Briggs, C. Bungaro, M. Buongiorno Nardelli, J.-L. Fattebert, K. Rapcewicz, C. Roland, W.G. Schmidt, and Q. Zhao, phys, stat. Sol. (b) **217**, 685 (2000)
[5] A. Edelman, T. A. Arias, and S. T. Smith, SIAM J. on Matrix Anal. Appl. **20**, 303 (1998)
[6] J.-L. Fattebert and J. Bernholc, Phys. Rev. B **62**, 1713 (2000).
[7] E.B. Stechel, A.R. Williams, and Peter J. Feibelman, Phys. Rev. B **49**, 10 088 (1994).
[8] V. A. Gubanov and C. Y. Fong, Appl. Phys Lett. **75**, 88 (1999).
[9] F. Mauri and G. Galli, Phys. Rev. B **50**, 4316 (1994).
[10] P. Ordejón, D. A. Drabold, R. M. Martin, et al., Phys. Rev. B **51**, 1456 (1995).
[11] D. Raczkowski, C.Y. Fong, and E.B. Stechel, submitted.
[12] J. Kim, F. Mauri, and G. Galli, Phys. Rev. B **52**, 1640 (1995).
[13] W. Hierse and E.B. Stechel, Phys. Rev. B **54**, 16 515 (1996).

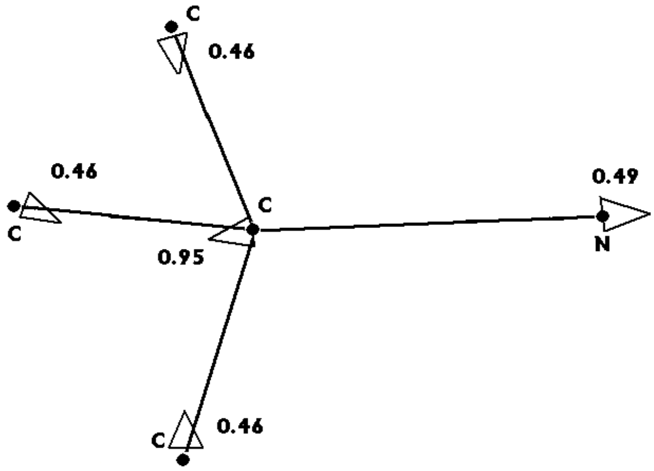

FIG. 2: Fig. 2 Movements from ideal cubic positions around N defect for configuration 2 (C2)

	E_{C1}-E_{C2} (Ry)
64 atoms	0.2833
216 atoms	0.2545

TABLE I: Table 1 Relative energy difference between the two configurations with diagonalization

	E_{C1}-E_{C2} (Ry)
S1	0.2661
S2	0.2146
S3	0.2837
S4	0.2822
S5	0.2805

TABLE II: Table 2. Relative energy difference between the two 64-atom configurations with different localization region.

	E_{C1}-E_{C2} (Ry)
S3	0.2534
S6	0.24963

TABLE III: Table 3 Relative energies differences for the 216-atom unit cell.

Nitrogen defect 64-atom (C2)	Energy (Ry)
S3 Relaxed	-632.5009
S3 Single energy at final positions	-632.4862

TABLE IV: Table 4. Energies at final positions with and without relaxation from an initial geometry.

Nitrogen defect (C2)	Energy (Ry)
S3 Final geometry after 1 inputs of Φ	-632.4974
S3 Final geometry after 2 inputs of Φ	-632.5019
S3 Final geometry after 10 inputs of Φ	-632.5084

TABLE V: Table 5. Energies at final positions with different starting values for Φ.