JOURNAL OF COMPUTATIONAL PHYSICS **124,** 271–285 (1996)
ARTICLE NO. 0059

A Comparative Study on Methods for Convergence Acceleration of Iterative Vector Sequences

V. EYERT*

Max-Planck-Institut für Festkörperforschung, Heisenbergstraße 1, D-70569 Stuttgart, Germany

Received November 22, 1994; revised July 25, 1995

We discuss several methods for accelerating the convergence of the iterative solution of nonlinear equation systems commonly in use and point to interrelations between them. In particular we investigate two of the most sophisticated schemes, namely the Anderson mixing and the Broyden update, both generalized to the consideration of arbitrarily many previous iterations. For the Broyden method we give a new derivation which is much simpler than that recently proposed by Vanderbilt and Louie. We show that if the additional parameters invented by these authors in order to increase flexibility are used to optimize the convergence of the iteration process they in fact cancel out. In addition we prove that in this (optimal) case the Anderson mixing and the Broyden update as applied to the inverse Jacobian are fully identical. Thus we come to the conclusion that neither of these schemes is superior. Moreover, we show that Broyden update of the inverse Jacobian is superior to updating the Jacobian itself. Finally we propose an extension of the Anderson mixing which avoids the numerical difficulties all these methods are faced with.

I. INTRODUCTION

Both the application and development of effective techniques for multidimensional optimization and the solution of nonlinear equation systems have been the subject of a rapidly growing interest in the last years [1–6]. The method of choice in this context is iteration and so we are likewise faced with an extensive work on schemes to achieve and accelerate convergence. According to common appraisal the most sophisticated ones nowadays seem to be conjugate-gradient, Newton–Raphson and quasi-Newton–Raphson, modification, or variable metric schemes.

Like in many other areas methods for improving convergence have also found entrance in the authors own research field, namely first principles electronic structure calculations of condensed matter. Here we rely on density functional theory and the local density approximation which in an approximate manner cast the full many-body problem connected with the system of interacting electrons into a single-particle self-consistent field problem [7, 8]. The (nonlinear) self-consistency equations are formulated in terms of the electronic charge density and after discretization they are solved by iteration (for a more detailed description see Ref. [9] and references therein).

Usually each of these iterations itself needs a lot of CPU time and thus reducing their number by accelerating the convergence would save computer resources or else allow for larger systems to be investigated [10]. As concerns these general demands, we believe electronic structure calculations are no exception.

In order to get a better feeling of which method would be best suited to the context just outlined we started the present analysis of convergence acceleration schemes. Here we concentrated especially on quasi-Newton–Raphson schemes and generalized secant methods. When applying the former class to nonlinear equation systems we are led to Broyden's rank-1 update [1, 3, 6, 11–15] and a recent generalization of this scheme which was initiated by Vanderbilt and Louie and improved the quality of the method substantially [16–18]. An example for the generalized secant methods, in contrast, is the Anderson mixing scheme which has been used quite often for self-consistent electronic structure calculations [1, 10, 19–22].

However, when going into more detail we realized that the actual formulation of the generalized Broyden method as given by Vanderbilt and Louie, as well as by Johnson, in fact, is too elaborate. This may be partially traced back to their inclusion of additional weights as a means to increase flexibility of the method. Still, as we observed, these weights actually can*not* be used to accelerate the iteration process. Hence, with the weights omitted we were able to derive a new formulation of a generalized Broyden method which is fully in the spirit of the original ansatz by Broyden and, thus, very simple.

The Broyden scheme can be used to update either an approximate Jacobian or else an approximate inverse Jacobian. In the language of the original Broyden update these two are referred to as the first and second methods, respectively [11]. As concerns the relation between these schemes, we came to the conclusion that updating the inverse Jacobian leads to faster convergence for principal reasons.

*Present address: Hahn-Meitner-Institut, Glienicker Straße 100, D-14109 Berlin; Email: eyert@physik.uni-augsburg.de.

0021-9991/96 $18.00

When dealing with the Broyden scheme for the inverse Jacobian we furthermore became aware of its full identity to the Anderson mixing. At the same time we were faced with the common opinion regarding the former that, in general, it is superior to the latter. Since, due to our investigations, this point of view is *not* justified and, moreover, might hinder the optimal use of the methods available, we saw the need for a few words which might help clarify the situation a bit.

We start out in the following section with a definition of the underlying mathematical problem and the corresponding notation. In addition, we present a test case which will be used to illustrate the findings of the following sections. We proceed with a survey of different methods, including the simple mixing, the Anderson mixing, and our new derivation of the generalized Broyden method. In Section VI we discuss the modified Broyden method by Vanderbilt and Louie, i.e. the generalized Broyden method with the weights included.

Section VII relates the different versions of the Broyden method to each other, and furthermore, turns to the detailed comparison of the Anderson and the Broyden scheme. In Section VIII we deal with numerical problems and propose an extension of the Anderson mixing. The conclusion finally summarizes the most important of our findings.

II. DEFINITION OF THE PROBLEM

In general all multidimensional iterative procedures may be defined by a nonlinear operator which when applied to an input vector $|x^{(l)}\rangle$ of length N produces an output vector $|y^{(l)}\rangle$ of the same length. Here the superscript l denotes the iteration number, $l \geq 1$ and we have used Dirac's notation. Self-consistency is achieved when both vectors coincide or, equivalently, when the residual vector defined by

$$|F^{(l)}\rangle := |y^{(l)}\rangle - |x^{(l)}\rangle \tag{2.1}$$

vanishes.

Since it is actually the length of the residual vector which will be used as a measure of the convergence reached so far it is useful to define a norm in the space spanned by the input and output vectors. Going on we define a scalar product of two vectors $|a\rangle$ and $|b\rangle$ by

$$\langle a|b\rangle := \sum_{i,j=1}^{N} a_i g_{ij} b_{j.} \tag{2.2}$$

Here we have, in addition, invented a positive definite metric which is useful when dealing with inhomogeneous iteration vectors. An example often quoted in this context is the simultaneous iteration of charge and spin densities [22].

In order to achieve convergence, i.e., to arrive at a vanishing norm $\langle F^{(l)}|F^{(l)}\rangle = 0$ as soon as possible one has to combine all the vectors known from previous iterations in an optimal way in order to provide for the most promising input vector for the following iteration, $|x^{(l+1)}\rangle$. As already mentioned in the Introduction different schemes have been developed for approaching this main goal in the past. Some of these we will outline next in order of increasing complexity. According to the context of electronic structure calculations we will thereby limit ourselves to the case where the exact Jacobian

$$J_{ij}^{(l)} = \frac{\partial F_i^{(l)}}{\partial x_j^{(l)}} \tag{2.3}$$

is not available. Still we avoid using the underlying physics by incorporating response functions [13, 14].

Throughout this work we will use a simple numerical example to illustrate our results. This test case was proposed by Vanderbilt and Louie and is defined by the following recipe to calculate the residual vector for a given input vector [16]:

$$F_i = -d_{ii}x_i - cx_i^3 \quad \text{for } i = 1, ..., 5. \tag{2.4}$$

Obviously the solution vector has the components $x_i = 0$. The diagonal matrix d and the scalar c are given by

$$d_{ii} = (3.0, 2.0, 1.5, 1.0, 0.5), \quad c = 0.01 \tag{2.5}$$

and the starting vector is chosen to have the components

$$x_i = 1. \tag{2.6}$$

In contrast to many more realistic test calculations the above example has the distinct advantage that it could be very easily implemented, as well as reproduced.

III. SIMPLE MIXING

In the simple mixing scheme the input and the output vector of the actual iteration are just mixed linearly this resulting in the following input vector for the forthcoming iteration:

$$|x^{(l+1)}\rangle = |x^{(l)}\rangle + \beta^{(l)}|F^{(l)}\rangle, \tag{3.1}$$

where $\beta^{(l)}$ is the socalled mixing parameter which may vary from one iteration to another but usually does not. Obviously $\beta^{(l)} = 0$ means to take the old input vector once more and $\beta^{(l)} = 1$ corresponds to using the output vector of the actual iteration directly as input vector for the following iteration. The latter choice in most cases leads to oscillations and, if not damped, to a divergent iteration process. In general the simple mixing depends very sensitively on the choice of the mixing parameter, but experience has shown that setting $\beta^{(l)} \approx 0.5$ or a little less is a good choice for simple systems. However, as Dederichs and Zeller have outlined in a detailed study there exist many cases where $\beta^{(l)}$ has to be fixed to a much smaller value [20].

The simple mixing usually needs many iterations before convergence is reached since it forgets about all the information gained from previous iterations and, even worse, fixes the linear combination of input and output vector arbitrarily.

IV. ANDERSON MIXING

Whereas in the simple mixing only the input and output vector of the actual iteration are used to calculate the new input vector the mixing scheme introduced by Anderson, in addition, takes the vectors of the previous iterations into account and combines them in a much better way [19]. Though the Anderson scheme is one of the most powerful methods which leads to very fast convergence, as we will see later on, it is still quite simple as concerns the underlying concept.

Let us begin its description by first defining general vectors in the M-dimensional subspaces spanned by the input and output vectors, respectively, of the M previous iterations to be considered. They are given by

$$|\bar{x}^{(l)}\rangle := |x^{(l)}\rangle + \sum_{j=1}^{M} \vartheta_j^{(l)} \left(|x^{(l-j)}\rangle - |x^{(l)}\rangle\right) \tag{4.1}$$

in the space spanned by the input vectors and by a corresponding equation for the output vectors. Note, that $0 \leq M \leq l - 1$ and that the coefficients $\vartheta_j^{(l)}$ are as yet not specified. Again we prefer working with the residual vectors instead of output vectors and define the general residual vector by

$$|\bar{F}^{(l)}\rangle := |F^{(l)}\rangle + \sum_{j=1}^{M} \vartheta_j^{(l)} \left(|F^{(l-j)}\rangle - |F^{(l)}\rangle\right). \tag{4.2}$$

Now we fix the coefficients $\vartheta_j^{(l)}$ by the requirement that they yield that particular linear combination which minimizes the norm of the general residual vector $E := \langle \bar{F}^{(l)} | \bar{F}^{(l)} \rangle$. Minimizing with respect to the coefficients $\vartheta_j^{(l)}$ we are led to the linear equation system

$$\sum_{j=1}^{M} \langle F^{(l)} - F^{(l-i)} | F^{(l)} - F^{(l-j)} \rangle \vartheta_j^{(l)} = \langle F^{(l)} - F^{(l-i)} | F^{(l)} \rangle \quad \forall_{i=1,\dots,M}, \tag{4.3}$$

which has to be solved in order to yield the coefficients $\vartheta_j^{(l)}$ and, hence, the linear combination with the shortest residual vector.

Having found this optimal linear combination we fix the input vector for the subsequent iteration by

$$|x^{(l+1)}\rangle = |\bar{x}^{(l)}\rangle + \beta^{(l)} |\bar{F}^{(l)}\rangle, \tag{4.4}$$

where $\beta^{(l)}$ again is the mixing parameter. This looks quite similar to the recipe used for the simple mixing. However, there we mixed the input and output vectors of the actual iteration directly, whereas now we use the optimal linear combination of the input and output vectors within the spaces spanned by the vectors of the M previous iterations. This means that in the Anderson mixing the memory of the whole iteration process is built in which helps finding the final solution quite fast. Still the Anderson mixing scheme may be reduced to the simple mixing by setting $M = 0$ thus switching the memory off.

It has been argued by several authors, including Anderson, that near convergence linear dependencies within the system of residual vectors might evolve which hinder the accurate solution of the above linear equation system [19, 17]. For this reason it was proposed to reduce the number M of previous iterations to be mixed in to only a few and, actually, in many applications M is set to two. As we will explain in more detail in Section VIII these arguments do not really hold since there exist very simple means to avoid problems of that kind. Nevertheless, as Anderson has pointed out the power of his method increases slowly for $M > 3$ since then we only mix in poor vectors of the early iterations. Hence he suggested that in practice $M = 4$ or 5 should be used.

As concerns the mixing parameter $\beta^{(l)}$ Anderson proposed to avoid the choice $\beta^{(l)} = 0$ in order not to get stuck in the space spanned by the previous input vectors. In contrast, he found $\beta^{(l)} = 1$ most appropriate but admitted that the optimal value should be adjusted empirically. In this context it should be noted that since we minimize the length of the vector $|\bar{F}^{(l)}\rangle$, the second term in Eq. (4.4) and, hence, the influence of the mixing parameter becomes smaller when approaching convergence.

Next we focus the readers attention to a particular assumption underlying the Anderson mixing, namely that the linear approximation holds when expanding a residual vector in terms of input vectors near the optimal position. This is implied by choosing the same coefficients in the linear combinations (4.1) and (4.2) for the general input and residual vector and should be kept in mind when comparing to Broyden's methods.

As concerns storage and computer time considerations the Anderson method needs storage of the (2 M + 2) N-component vectors $|x^{(l)}\rangle$ and $|F^{(l)}\rangle$ of the actual and the M previous iterations, and, furthermore the solution of a linear equation system of order M. This is quite moderate and seems to be the minimum to be reached by the more sophisticated methods.

V. BROYDEN UPDATE

We now turn to the discussion of the methods introduced by Broyden including the recent developments introduced by Vanderbilt and Louie which put the method much forward [11, 16, 17]. Nevertheless, as already mentioned in the introduction the new ideas of the latter authors actually may be incorporated in the original scheme in a much more straightforward way which leads to what we call the generalized Broyden method. This we will outline in the following whereas the discussion of the modified Broyden method worked out by Vanderbilt and Louie and its comparison to our formalism are postponed to Sections VI and VII.

In order to make the simplicity of our approach even more clear and to relate it to the original ansatz by Broyden we start out with a brief description of that scheme:

The approach underlying the methods introduced by Broyden at first glance seems to be quite different from the Anderson mixing scheme discussed before [1, 3, 6, 11, 12]. It does not aim at a minimization of a general residual vector, but rather at an optimal Jacobian from which an input vector for the following iteration can be calculated.

Broyden's first method is a quasi-Newton–Raphson method updating an approximate Jacobian. Like the Anderson mixing it starts out from the assumption that the linear approximation is justified when expanding the residual vector in terms of the input vectors and, hence, the residual vector of the subsequent iteration could be written as

$$|F^{(l+1)}\rangle \approx |F^{(l)}\rangle + J^{(l)}(|x^{l+1}\rangle - |x^{(l)}\rangle), \tag{5.1}$$

where $J^{(l)}$ is the Jacobian, approximated by

$$J_{ij}^{(l)} \approx \frac{F_i^{(l+1)} - F_i^{(l)}}{x_j^{(l+1)} - x_j^{(l)}}. \tag{5.2}$$

If, furthermore, the residual vector on the left-hand side of Eq. (5.1) is required to vanish we arrive at the following proposal for the new input vector:

$$|x^{(l+1)}\rangle = |x^{(l)}\rangle - (J^{(l)})^{-1}|F^{(l)}\rangle. \tag{5.3}$$

However, we do not really expect the linear approximation to work well far away from the final solution and, hence, the residual vector $|F^{(l+1)}\rangle$ in most cases does not vanish. Then we would have to find a new estimate for the Jacobian $J^{(l+1)}$. This we will fix by the condition that first it has to fulfil Eq. (5.1) above, namely,

$$|\Delta F^{(l)}\rangle = J^{(l+1)}|\Delta x^{(l)}\rangle, \tag{5.4}$$

which $J^{(l)}$ did not; here we have defined the difference vectors

$$|\Delta x^{(l)}\rangle := |x^{(l+1)}\rangle - |x^{(l)}\rangle \tag{5.5}$$

and

$$|\Delta F^{(l)}\rangle := |F^{(l+1)}\rangle - |F^{(l)}\rangle. \tag{5.6}$$

The important point to note now is that the constraint (5.4) fixes only the projection of the new Jacobian onto the vector $|\Delta x^{(l)}\rangle$. Clearly, the same holds for the update $J^{(l+1)} - J^{(l)}$ for which we have from Eq. (5.4)

$$(J^{(l+1)} - J^{(l)})|\Delta x^{(l)}\rangle = |\Delta F^{(l)}\rangle - J^{(l)}|\Delta x^{(l)}\rangle. \tag{5.7}$$

Being a bit formal we next write the update as

$$\begin{aligned} J^{(l+1)} - J^{(l)} &= (J^{(l+1)} - J^{(l)}) \frac{|\Delta x^{(l)}\rangle\langle\Delta x^{(l)}|}{\langle\Delta x^{(l)}|\Delta x^{(l)}\rangle} \\ &\quad + (J^{(l+1)} - J^{(l)}) \left[\mathcal{I} - \frac{|\Delta x^{(l)}\rangle\langle\Delta x^{(l)}|}{\langle\Delta x^{(l)}|\Delta x^{(l)}\rangle} \right], \end{aligned} \tag{5.8}$$

where $\mathcal{I}$ is the unity operator and the second term just presents the projection of the update onto the space orthogonal to $|\Delta x^{(l)}\rangle$. As Broyden argued, there is no reason for this projection not to vanish and this way he arrived at his rank-1 update formula,

$$\begin{aligned} J^{(l+1)} - J^{(l)} &= \frac{1}{\langle\Delta x^{(l)}|\Delta x^{(l)}\rangle} \\ &\quad (|\Delta F^{(l)}\rangle\langle\Delta x^{(l)}| - J^{(l)}|\Delta x^{(l)}\rangle\langle\Delta x^{(l)}|). \end{aligned} \tag{5.9}$$

As an initial guess for the Jacobian $J^{(1)}$, Broyden proposed to use a constant diagonal matrix corresponding to a simple mixing for the first iteration, i.e.,

$$J^{(1)} = -\frac{1}{\beta^{(1)}} \mathcal{I}. \tag{5.10}$$

Later on Dennis and Moré pointed out that the Broyden update formula (5.9) could be equally well derived if the Jacobian $J^{(l+1)}$ is forced to fulfil Eq. (5.4) and, in addition, the distance between the Jacobian and its predecessor $J^{(l)}$ is required to be minimal in the sense of the Frobenius norm [12]. In other words, we have to minimize the Frobenius norm

$$E = \|J^{(l+1)} - J^{(l)}\|^2 = \sum_{ij} |J_{ij}^{(l+1)} - J_{ij}^{(l)}|^2 \tag{5.11}$$

with the constraint (5.4). Defining

$$E' = E + \langle\lambda|[|\Delta F^{(l)}\rangle - J^{(l+1)}|\Delta x^{(l)}\rangle], \tag{5.12}$$

where $|\lambda\rangle$ is a vector of Lagrange multipliers, we minimize above expression with respect to the elements of the Jacobian $J^{(l+1)}$ and arrive at the rank-1 update

$$J^{(l+1)} - J^{(l)} = \tfrac{1}{2}|\lambda\rangle\langle\Delta x^{(l)}|. \tag{5.13}$$

Inserting this into Eq. (5.4) we get

$$|\Delta F^{(l)}\rangle - J^{(l)}|\Delta x^{(l)}\rangle - \tfrac{1}{2}|\lambda\rangle\langle\Delta x^{(l)}|\Delta x^{(l)}\rangle = 0 \tag{5.14}$$

which when combined with Eq. (5.13) again leads to the Broyden update formula (5.9). Hence the derivation based on a minimization is tailored such that it reproduces the original result [12]. This should be kept in mind whenever generalizing the Broyden method and using the formulation as a minimization process.

The Broyden update in the form (5.9) still has two distinct disadvantages: It needs the storage of the full $N \times N$ Jacobian and in order to use it for Eq. (5.3) we have to invert this, usually huge, matrix.

However, as is well known, the latter problem may be easily circumvented by employing Broyden's second method which is based on updating the inverse Jacobian instead of the Jacobian itself. The condition (5.4) for the inverse Jacobian $G^{(l+1)}$ to be fulfilled then reads as

$$|\Delta x^{(l)}\rangle = G^{(l+1)}|\Delta F^{(l)}\rangle. \tag{5.15}$$

Everything else proceeds as above and finally leads to the following update formula for the inverse Jacobian,

$$G^{(l+1)} - G^{(l)} = \frac{1}{a_{ll}}(|\Delta x^{(l)}\rangle\langle\Delta F^{(l)}| - G^{(l)}|\Delta F^{(l)}\rangle\langle\Delta F^{(l)}|). \tag{5.16}$$

Here the matrix a is the overlap matrix of the difference residual vectors, i.e.,

$$a_{nk} := \langle\Delta F^{(n)}|\Delta F^{(k)}\rangle. \tag{5.17}$$

Again we use for the first iteration the simple mixing according to

$$G^{(1)} = -\beta^{(1)}\mathscr{I}. \tag{5.18}$$

Broyden's second method thus avoids the inversion of the Jacobian without loss of accuracy, but still we have to store the huge matrix. Leaving this problem for later, we will first turn to the above-mentioned generalization of the Broyden update which is based on an important argument made by Vanderbilt and Louie [16]. They started out from the observation that in the original approach taken by Broyden the Jacobian is updated from the knowledge of the actual iteration only and the information gained by all previous iterations is not explicitly taken into account. As they pointed out, this way the memory as contained in the Jacobian may be arbitrarily overwritten and, hence, the systematics of the whole scheme is lost.

In order to overcome this weakness Vanderbilt and Louie claimed that the forthcoming Jacobian $J^{(l+1)}$ should not only satisfy Eq. (5.4) but, in general, the set of conditions

$$|\Delta F^{(m)}\rangle = J^{(l+1)}|\Delta x^{(m)}\rangle \quad \forall_{m=1,\ldots,l}. \tag{5.19}$$

Due to this ansatz all the information collected so far in the iteration process could be used in an optimal way.

Taking this as a starting point Vanderbilt and Louie derived a modified Broyden method which generalizes the original scheme. However, unfortunately they formulated their ideas on the basis of Broyden's first method; i.e., they worked with the Jacobian itself which has the above-mentioned disadvantage.

Moreover, and of even higher importance, when generalizing the Broyden method for the inclusion of more than one condition, Vanderbilt and Louie, in addition, introduced a set of weights in order to gain more flexibility of the scheme, resulting in the so-called modified Broyden method. However, as we will show in detail in Section VII, if these weights are interpreted as free parameters and adjusted to optimal convergence of the iteration process they in fact cancel out. Hence, they could be omitted from the very beginning and the whole formalism would be much simplified. For this reason we will not follow these authors here but, instead, derive a rather straightforward generalization of the Broyden update which, in opinion, is the simplest possible and which has the additional advantage of being fully in the spirit of the original scheme by Broyden. Thereby we will concentrate on Broyden's second method, i.e., the update of the inverse Jacobian.

So we start out writing down the set of conditions for the inverse Jacobian $G^{(l+1)}$ which corresponds to Eq. (5.19) and generalizes Eq. (5.15), i.e.,

$$|\Delta x^{(m)}\rangle = G^{(l+1)}|\Delta F^{(m)}\rangle \quad \forall_{m=l+1-M,\ldots,l}. \tag{5.20}$$

In contrast to the constraint (5.19) given by Vanderbilt and Louie we here have limited the number of conditions to M previous iterations. This way we are able to get a better control over the method, especially when comparing it to other schemes. Apparently, by setting $M = l$ we get back Vanderbilt and Louie's proposal.

An important point to note now is that the step taken by Vanderbilt and Louie, namely the inclusion of the whole set of conditions (5.20) changes the approach underlying the Broyden method completely. Whereas in the original scheme each new condition of the type (5.15) was used to update the inverse Jacobian from its direct predecessor the method now ignores all the inverse Jacobians calculated before and, instead, uses the conditions (5.20) and the initial guess to evaluate the new inverse Jacobian directly. This, furthermore, implies that, strictly speaking, the method by now is no update scheme any more but, rather, a direct optimization of the inverse Jacobian. This becomes most obvious when setting $M = l$, hence, when including all previous iterations.

With this in mind we now rewrite Eq. (5.20) as

$$(G^{(l+1)} - G^{(l+1-M)})|\Delta F^{(m)}\rangle = |\Delta x^{(m)}\rangle - G^{(l+1-M)}|\Delta F^{(m)}\rangle \quad \forall_{m=l+1-M,\dots,l}, \tag{5.21}$$

which is just the set of conditions for the update $G^{(l+1)} - G^{(l+1-M)}$. As above, we present the latter in the form

$$G^{(l+1)} - G^{(l+1-M)} = (G^{(l+1)} - G^{(l+1-M)})\,\mathscr{P}_F + (G^{(l+1)} - G^{(l+1-M)})(\mathscr{I} - \mathscr{P}_F) \tag{5.22}$$

where $\mathscr{P}_F$ is the projection operator onto the space spanned by the states $|\Delta F^{(m)}\rangle$, $m = l + 1 - M, \dots, l$. This operator is defined in the usual manner as

$$\mathscr{P}_F := \sum_{n,k=l+1-M}^{l} (a^{-1})_{nk}|\Delta F^{(n)}\rangle\langle\Delta F^{(k)}|, \tag{5.23}$$

where the matrix a is the overlap matrix (5.17).

Next we follow again Broyden's argument that the projection of the inverse Jacobian onto the space orthogonal to that spanned by the states $|\Delta F^{(m)}\rangle$ and, hence, the second term in Eq. (5.22) should vanish. Combining (5.21) and (5.22) we thus arrive at the desired rank-M update formula:

$$G^{(l+1)} - G^{(l+1-M)} = \sum_{m,k=l+1-M}^{l} (a^{-1})_{mk}[|\Delta x^{(m)}\rangle\langle\Delta F^{(k)}| - G^{(l+1-M)}|\Delta F^{(m)}\rangle\langle\Delta F^{(k)}|] \tag{5.24}$$

which is the proper generalization of the rank-1 update formula (5.16) of the original Broyden method and reduces exactly to that result for $M = 1$. As an initial guess for the inverse Jacobian we finally use

$$G^{(l+1-M)} = -\beta^{(l+1-M)}\,\mathscr{I}. \tag{5.25}$$

In order to prepare for the later discussion of the approach due to Vanderbilt and Louie we complement the previous derivation with the corresponding one, based on a constrained minimization of the Frobenius norm. In the present context this means that we have to minimize

$$E' = \|G^{(l+1)} - G^{(l+1-M)}\|^2 + \sum_{m=l+1-M}^{l} \langle\lambda^{(m)}|[|\Delta x^{(m)}\rangle - G^{(l+1)}|\Delta F^{(m)}\rangle], \tag{5.26}$$

where the second term traces back to the set of constraints (5.20) which are coupled by vectors $|\lambda^{(m)}\rangle$ of Lagrange multipliers. Minimizing E' with respect to the matrix elements of the inverse Jacobian $G^{(l+1)}$ we get the rank-M update

$$G^{(l+1)} - G^{(l+1-M)} = \tfrac{1}{2}\sum_{k=l+1-M}^{l} |\lambda^{(k)}\rangle\langle\Delta F^{(k)}| \tag{5.27}$$

which when inserted into Eq. (5.20) yields

$$|\Delta x^{(m)}\rangle - G^{(l+1-M)}|\Delta F^{(m)}\rangle - \tfrac{1}{2}\sum_{k=l+1-M}^{l} |\lambda^{(k)}\rangle\langle\Delta F^{(k)}|\Delta F^{(m)}\rangle = 0 \quad \forall_{m=l+1-M,\dots,l}. \tag{5.28}$$

Multiplying this with the inverse of the overlap matrix, we get

$$\tfrac{1}{2}|\lambda^{(k)}\rangle = \sum_{m=l+1-M}^{l} (a^{-1})_{mk}[|\Delta x^{(m)}\rangle - G^{(l+1-M)}|\Delta F^{(m)}\rangle] \tag{5.29}$$

and, together with Eq. (5.27), we finally again arrive at the rank-M update formula (5.24).

To sum up, we combine Eq. (5.24) with Eq. (5.25) and, in order to get the input vector of the following iteration, we finally insert it into Eq. (5.3), resulting in

$$|x^{(l+1)}\rangle = |x^{(l)}\rangle - G^{(l-M)}|F^{(l)}\rangle - \sum_{m=l-M}^{l-1} \gamma_m^{(l)}[|\Delta x^{(m)}\rangle - G^{(l-M)}|\Delta F^{(m)}\rangle], \tag{5.30}$$

where we have abbreviated

$$\gamma_m^{(l)} = \sum_{k=l-M}^{l-1} (a^{-1})_{mk}\langle\Delta F^{(k)}|F^{(l)}\rangle. \tag{5.31}$$

From this last equation it becomes obvious that even the explicit inversion of the overlap matrix a can be circumvented since Eq. (5.31) presents just a linear equation system with solution vector $\gamma^{(l)}$ [17, 18].

Furthermore, we point out that also the above-mentioned problem of storing the huge inverse Jacobian matrix has been resolved by now. This is, first, due to the calculation of the inverse Jacobian by direct optimization instead of stepwise update, as explained before. Second, we here take advantage of the observation made by Srivastava that the update as well as the inverse Jacobian itself is just a sum of dyadic products which enter the formalism only when applied to residual vectors as in Eq. (5.3) [14]. Hence by combining the vectors in these matrix–vector products just the other way round we arrive at products of dot-products and vectors and all the huge matrices have collapsed.

Altogether, we are left with storing the two N-component vectors $|x^{(l)}\rangle$ and $|F^{(l)}\rangle$ of the actual and the $2M$ vectors $|\Delta x^{(m)}\rangle$ and $|\Delta F^{(m)}\rangle$ of the M previous iterations. In addition we have to solve a linear equation system of order M.

VI. MODIFIED BROYDEN METHOD

In order to allow for a thorough comparison of the methods discussed in this work which will be aimed at in Section VII we now give a brief sketch of the modified Broyden method proposed by Vanderbilt and Louie [16]. However, for the reasons discussed previously we here turn directly to its formulation in terms of the inverse Jacobian.

As already mentioned before, Vanderbilt and Louie started out from their observation that the Jacobian to be updated should fulfil the conditions (5.19) all at once. In the modified Broyden method founded by these authors this is achieved by minimization of a functional which contains the Frobenius norm, as well as all the conditions (5.19), and which in terms of the inverse Jacobian reads as [16, 17]

$$\begin{aligned} E = {} & (w_0^{(l+1)})^2 \| G^{(l+1)} - G^{(l+1-M)} \|^2 \\ & + \sum_{m=l+1-M}^{l} \frac{(w_m^{(l+1)})^2}{a_{mm}} \, | \, |\Delta x^{(m)}\rangle - G^{(l+1)} |\Delta F^{(m)}\rangle |^2. \end{aligned} \tag{6.1}$$

(Note, that our definition of the inverse Jacobian differs by a minus sign from that of above authors).

As concerns the functional (6.1) the differences to our approach (5.26) are easily stated: First, instead of strictly enforcing the constraints (5.20) by coupling them to the minimization of the Frobenius norm via Lagrange multipliers they are here incorporated directly as part of the functional to be minimized. Second, as a consequence and in contrast to our approach the inverse Jacobian here enters each term in the sum quadratically which will lead to the inversion of a huge $N \times N$ matrix, even if this can be handled by applying the Sherman–Morrison formula [1] or else be reduced to the inversion of a $M \times M$ matrix as we will see later on. Third, there now appear the already mentioned weights $w_m^{(l+1)}$ which were introduced by Vanderbilt and Louie in order to provide for a greater flexibility of the method and which may be fixed each iteration anew (hence the superscript).

Minimizing now the functional E with respect to the matrix elements of the inverse Jacobian we get

$$\begin{aligned} (w_0^{(l+1)})^2 (G^{(l+1)} - G^{(l+1-M)}) = {} & \sum_{m=l+1-M}^{l} \frac{(w_m^{(l+1)})^2}{a_{mm}} \\ & [|\Delta x^{(m)}\rangle\langle\Delta F^{(m)}| - G^{(l+1)} |\Delta F^{(m)}\rangle\langle\Delta F^{(m)}|]. \end{aligned} \tag{6.2}$$

Note, that since the inverse Jacobian $G^{(l+1)}$ entered Eq. (6.1) quadratically it here still appears on the right-hand side within the sum. Solving for $G^{(l+1)}$ leads to

$$G^{(l+1)} = A^{(l+1)} [B^{(l+1)}]^{-1}, \tag{6.3}$$

where

$$A^{(l+1)} = (w_0^{(l+1)})^2 G^{(l+1-M)} + \sum_{m=l+1-M}^{l} \frac{(w_m^{(l+1)})^2}{a_{mm}} |\Delta x^{(m)}\rangle\langle\Delta F^{(m)}| \tag{6.4}$$

and

$$B^{(l+1)} = (w_0^{(l+1)})^2 \, \mathscr{I} + \sum_{m=l+1-M}^{l} \frac{(w_m^{(l+1)})^2}{a_{mm}} |\Delta F^{(m)}\rangle\langle\Delta F^{(m)}|. \tag{6.5}$$

At first glance it seems that by now we are faced with the inversion of a large $(N \times N)$-matrix since the vectors $|\Delta F^{(m)}\rangle$ entering $B^{(l+1)}$ are of length N. However, as can be read off from Eq. (6.5), $B^{(l+1)}$ simply consists of a linear combination of the unity operator and M projection operators onto M one-dimensional but mutually not necessarily orthogonal subspaces, where usually M is much smaller than N. As a consequence the inversion of $B^{(l+1)}$ may be reduced to the inversion of an $(M \times M)$-matrix. In short, this procedure is equivalent to calculating the resolvent of an operator given in terms of its eigenvalues and the projection operators onto the corresponding eigenspaces.

Being a bit more specific, we first rewrite the operator $B^{(l+1)}$ as

$$B^{(l+1)} = (w_0^{(l+1)})^2 (\mathscr{I} - \mathscr{P}_F) + \mathscr{P}_F B^{(l+1)} \mathscr{P}_F, \tag{6.6}$$

where $\mathscr{P}_F$ is the projector onto the space spanned by the states $(w_m^{(l+1)}/\sqrt{a_{mm}}) \, |\Delta F^{(m)}\rangle$, $m = l + 1 - M, ..., l$. From this the inverse may be written down immediately as

$$[B^{(l+1)}]^{-1} = \frac{1}{(w_0^{(l+1)})^2} (\mathscr{I} - \mathscr{P}_F) + \mathscr{P}_F [B^{(l+1)}]^{-1} \mathscr{P}_F, \tag{6.7}$$

where, however, taking the inverse of $B^{(l+1)}$ on the right-hand side only poses an $(M \times M)$-problem. Leaving the details of this evaluation to the Appendix we note the final result,

$$[B^{(l+1)}]^{-1} = \frac{1}{(w_0^{(l+1)})^2} \left[\mathscr{I} - \sum_{k,n=l+1-M}^{l} \frac{w_n^{(l+1)}}{\sqrt{a_{nn}}} \frac{w_k^{(l+1)}}{\sqrt{a_{kk}}} ((\tilde{b}^{(l+1)})^{-1})_{nk} |\Delta F^{(n)}\rangle\langle\Delta F^{(k)}| \right], \tag{6.8}$$

where the matrix $\tilde{b}^{(l+1)}$ is defined by

$$\tilde{b}_{nk}^{(l+1)} := (w_0^{(l+1)})^2 \delta_{nk} + \tilde{a}_{nk}^{(l+1)} \tag{6.9}$$

and

$$\tilde{a}_{nk}^{(l+1)} := \frac{w_n^{(l+1)}}{\sqrt{a_{nn}}} \frac{w_k^{(l+1)}}{\sqrt{a_{kk}}} \langle\Delta F^{(n)}|\Delta F^{(k)}\rangle. \tag{6.10}$$

Combining now Eqs. (6.3), (6.4), and (6.8) we get for the update of the inverse Jacobian

$$G^{(l+1)} - G^{(l+1-M)} = -G^{(l+1-M)} \sum_{k,n=l+1-M}^{l} \frac{w_n^{(l+1)}}{\sqrt{a_{nn}}} \frac{w_k^{(l+1)}}{\sqrt{a_{kk}}} ((\tilde{b}^{(l+1)})^{-1})_{nk} |\Delta F^{(n)}\rangle\langle\Delta F^{(k)}| + \frac{1}{(w_0^{(l+1)})^2} \sum_{m,k=l+1-M}^{l} \frac{w_m^{(l+1)}}{\sqrt{a_{mm}}} \frac{w_k^{(l+1)}}{\sqrt{a_{kk}}} \left(\delta_{mk} - \sum_{n=l+1-M}^{l} \tilde{a}_{mn}^{(l+1)} ((\tilde{b}^{(l+1)})^{-1})_{nk} \right) \times |\Delta x^{(m)}\rangle\langle\Delta F^{(k)}|. \tag{6.11}$$

Simplifying the brackets on the right-hand side with the help of Eq. (6.9) we finally arrive at the rank-M update formula

$$G^{(l+1)} - G^{(l+1-M)} = \sum_{k,n=l+1-M}^{l} \frac{w_n^{(l+1)}}{\sqrt{a_{nn}}} \frac{w_k^{(l+1)}}{\sqrt{a_{kk}}} ((\tilde{b}^{(l+1)})^{-1})_{nk} [|\Delta x^{(n)}\rangle\langle\Delta F^{(k)}| - G^{(l+1-M)} |\Delta F^{(n)}\rangle\langle\Delta F^{(k)}|] \tag{6.12}$$

which is the same as Eq. (5.24) of the generalized Broyden method with the elements of the inverse overlap matrix, $(a^{-1})_{nk}$ replaced by $w_n^{(l+1)} w_k^{(l+1)} ((\tilde{b}^{(l+1)})^{-1})_{nk}$. As a consequence when combining Eq. (6.12) with Eq. (5.3) in order to calculate the input vector for the following iteration we arrive at Eq. (5.30) with the just-mentioned replacement. Moreover, we can likewise avoid the explicit inversion of matrix $\tilde{b}$ and are left with the linear equation system

$$\sum_{m=l-M}^{l-1} \gamma_m^{(l)} \frac{\sqrt{a_{kk}}}{w_k^{(l+1)}} \frac{\sqrt{a_{mm}}}{w_m^{(l+1)}} \tilde{b}_{mk}^{(l)} = \langle\Delta F^{(k)}|F^{(l)}\rangle \tag{6.13}$$

to be solved for the coefficients $\gamma_m^{(l)}$.

In concluding this section we point to a different derivation of the modified Broyden method for the inverse Jacobian which was given by Johnson [17, 18]. It deviates from our Eq. (6.1) only in using $G^{(l)}$, instead of $G^{(l+1-M)}$. However, as we have pointed out in Section V, inserting $G^{(l)}$ here is not at all necessary when taking M conditions into account which contain all the information gained after determination of $G^{(l+1-M)}$. For this reason, use of $G^{(l)}$ does not really improve the scheme but, in contrast, requires in addition an inductive calculation of the inverse Jacobian, as well as the explicit inversion of the matrix $\tilde{b}$ (see Refs. [17, 18] for details).

VII. COMPARISON OF METHODS

After the description of the Anderson mixing, as well as several formulations of the Broyden update, we now turn to the comparison of these schemes. To this end we first relate the modified Broyden method as described in Section VI to the generalized scheme proposed in Section V:

As has become obvious from the previous derivations both schemes, despite their different starting points, arrive at results which differ only in that the modified Broyden method, in addition, contains the weights which were introduced originally by Vanderbilt and Louie in order to provide for a greater flexibility of the method.

Although such an ansatz indeed appears to be very promising, we will now show that this flexibility can*not* be used to accelerate the convergence of the iteration process. This is due to the following reasons:

Of the $M + 1$ conditions entering (6.1), those contained in the sum, i.e., those taken from Eq. (5.20), may be fulfilled independently since they fix only projections of the inverse Jacobian $G^{(l+1)}$ onto the M linear independent directions $|\Delta F^{(m)}\rangle$, $m = l + 1 - M, ..., l$. (For the time being we exclude the case of linear dependencies in the residual vector space which we will address in Section VIII.) As a consequence, when setting $w_0 = 0$ all the w_m-terms in Eq. (6.1) may be minimized independently with respect to the respective elements of the inverse Jacobian and, hence, all the corresponding weights w_m lose their meaning.

For $w_0 \neq 0$ the situation is a bit different since then, on every element of the inverse Jacobian which is subject to one of the w_m-conditions (5.20) actually, we impose a second condition mediated by the w_0-term. These two conditions can*not* be fulfilled independently but, rather, are in concurrence with the relative weights $w_0^2/(w_0^2 + w_m^2)$ and $w_m^2/(w_0^2 + w_m^2)$, respectively. The important point to notice now is that these two conditions push the inverse Jacobian

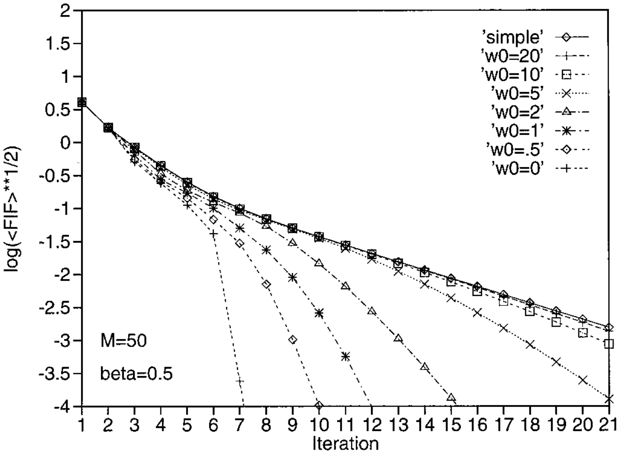

FIG. 1. Rate of convergence using the modified Broyden method for updating the inverse Jacobian with different values for w_0 (the curve marked $w_0 = 0$ actually was calculated with $w_0 = 0.01$ in order to avoid numerical instabilities). All other weights have been set to $w_m = 1$ for $m \geq 1$. The mixing parameter is $\beta = 0.5$.

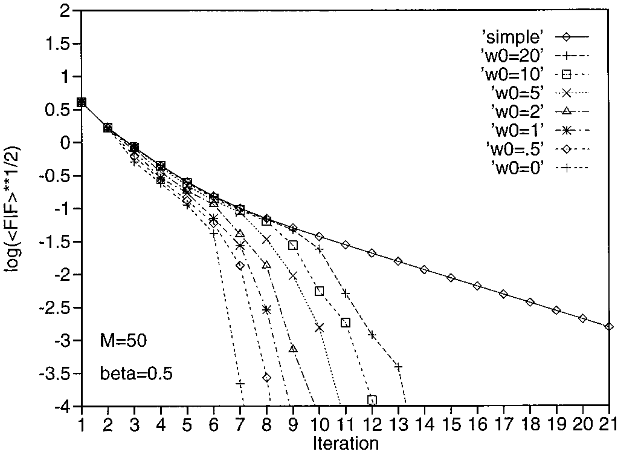

FIG. 3. Same as Fig. 1 but with weights $w_m = \langle F^{(m)}|F^{(m)}\rangle^{-1/2}$ for $m \geq 1$. Again to avoid instabilities the w_m are limited to the range $[1, 10^{12}]$.

to opposite directions: Whereas the w_m-conditions enforce an update of the inverse Jacobian according to the information gained from previous iterations the w_0-term in contrast hinders any changes of this matrix by requiring the Frobenius norm of the update to be minimal. This conflict becomes more obvious when investigating the situation $w_m^2 \ll w_0^2$ which means to ignore all the conditions (5.20) and to stay with the initial inverse Jacobian, hence, with simple mixing all the time.

As a consequence, in the intermediate range of weights we are just faced with a concurrence of simple mixing versus a full update and for any $w_0^2 > 0$ this results in an effective slowdown of the iteration process.

In order to demonstrate this we have implemented the full scheme proposed by Vanderbilt and Louie in the form outlined in the previous section and applied it to the test case presented in Section II. The results are shown in Figs. 1 to 4, where we have allowed for all previous iterations to be mixed in and have used two different values of the mixing parameter β. In each figure the weight w_0 varied from 0.01 to 20, whereas all other weights were fixed to $w_m = 1$ in Figs. 1 and 2 and to $w_m = 1/\langle F^{(m)}|F^{(m)}\rangle^{1/2}$ in Figs. 3 and 4. The latter choice was proposed by Johnson [17].

The curves presented support the arguments just given as they show the whole variation from full update to simple mixing when w_0 is increased. Even in Figs. 3 and 4, where we emphasize the "good" iteration vectors lying near the final solution by attaching "quality dependent" weights which reduce the relative influence of w_0, the curves for $w_0 > 0$ clearly deviate from the optimal result, $w_0 = 0$.

Furthermore, note the increasing sensitivity of the curves

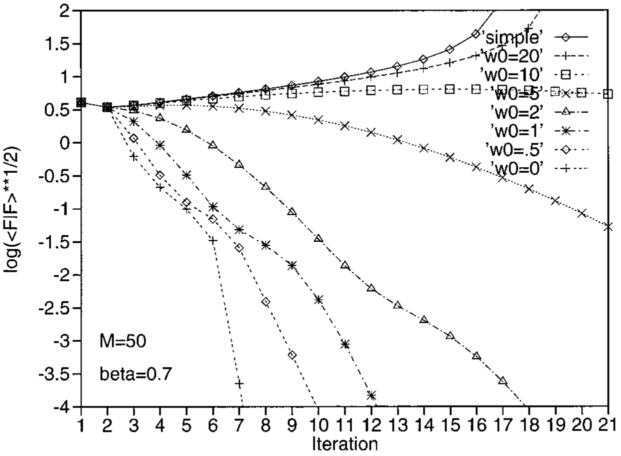

FIG. 2. Same as Fig. 1 but for $\beta = 0.7$.

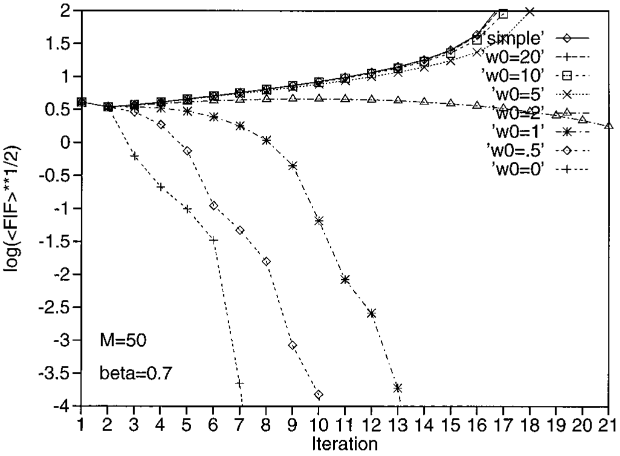

FIG. 4. Same as Fig. 3 but for $\beta = 0.7$.

to the mixing parameter β for the higher w_0-values, which is typical for simple mixing.

For completeness we add that for $w_0 > 0$, but not too large, the modified Broyden method in the version derived by Johnson yields slightly better results than the curves shown here. This traces back to the fact that his ansatz corresponds to Eq. (6.1) with $G^{(l+1-M)}$ replaced by $G^{(l)}$. As a consequence for values of w_0 of order one in Johnson's scheme the concurrence is not exactly between full update and simple mixing but, rather, between rank-M and rank-1 updates. Since the latter is better than simple mixing Johnson gets a bit faster convergence for w_0 lying in this medium range. For $w_0 \ll 1$ and for large w_0, however, this rank-1 property is suppressed and Johnson's method becomes more similar to the scheme by Vanderbilt and Louie.

Nevertheless, as Vanderbilt and Louie have claimed and as we have discussed in detail in Section V, exchanging the rank-M update with a rank-1 update just takes away the systematics of the whole scheme and for this reason, also, Johnson's method is faced with an effective deceleration for $w_0 > 0$. Finally, for $w_0 = 0$ the results of his scheme are identical to those shown here. Hence, for all versions of the modified Broyden method we arrive at the conclusion that when aiming at a fast convergence of the iteration process the best choice is to put $w_0 = 0$ in the ansatz (6.1).

However, as already argued above in this (optimal) case, also, all other weights w_m cancel out from the modified Broyden scheme. This becomes even more obvious on inspection of Eq. (6.13) which may be readily simplified to

$$\sum_{m=l-M}^{l-1} \gamma_m^{(l)} \left[\frac{(w_0^{(l+1)})^2}{(w_k^{(l+1)})^2} \delta_{mk} + 1 \right] a_{mk} = \langle \Delta F^{(k)} | F^{(l)} \rangle. \qquad (7.1)$$

The same result follows directly when looking at Eqs. (13) and (15) of Johnson or Eqs. (40) to (43) of van Leuken, who concentrated in particular on $w_0 \approx 0$ in their final equations [17, 18].

Now with all the weights taken away from the modified Broyden method the versions of Vanderbilt and Louie (in the form described in Section VII, i.e., as applied to the inverse Jacobian) as well as Johnson's become identical. Moreover, they arrive at exactly the same results as the generalized Broyden method formulated in Section VI.

In summary, if all the weights of the modified Broyden method are used to tune the iteration process to its fastest convergence they, in fact, cancel out and the results of this scheme, despite its different starting point (6.1), become identical to those of our much simpler formulation.

Although we have so far concentrated on the discussion of Broyden's method for the inverse Jacobian it is obvious that all the previous arguments likewise apply to the corresponding Broyden method for the Jacobian itself. In other words, the scheme presented by Vanderbilt and Louie, if tuned to optimal convergence, would lead to our generalized Broyden method as described in Section V, but formulated in terms of the Jacobian, instead of the inverse Jacobian. Hence, when investigating the relationship between Broyden's first and second method we can likewise work with the generalized method given in Section V and its counterpart based on the Jacobian itself.

At first glance there seems to be no reason why Broyden's first and second method should not perform equally well. Yet, as we will soon realize, Broyden's first method has a distinct disadvantage which could decelerate the iteration process. In order to demonstrate this we start out looking in more detail at the inverse Jacobian. Combining Eqs. (5.22) to (5.24) we get the identities

$$\begin{aligned} G^{(l+1)} &= G^{(l+1-M)}(\mathcal{I} - \mathcal{P}_F) \\ &\quad + \sum_{m,k=l+1-M}^{l} (a^{-1})_{mk} |\Delta x^{(m)}\rangle\langle \Delta F^{(k)}| \\ &= G^{(l+1-M)}(\mathcal{I} - \mathcal{P}_F) \qquad (7.2) \\ &\quad + \mathcal{P}_x \sum_{m,k=l+1-M}^{l} (a^{-1})_{mk} |\Delta x^{(m)}\rangle\langle \Delta F^{(k)}| \, \mathcal{P}_F \\ &= G^{(l+1-M)}(\mathcal{I} - \mathcal{P}_F) + \mathcal{P}_x G^{(l+1)} \mathcal{P}_F, \end{aligned}$$

where the projection operator $\mathcal{P}_x$ projects onto the space spanned by the states $|\Delta x^{(m)}\rangle$, $m = l + 1 - M, ..., l$ (we denote this as the x-space) and is defined in analogy with Eq. (5.23). A complementary relation to Eq. (7.2) holds for the Jacobian $J^{(l+1)}$.

At this point it should be noted that, in general, $J^{(l+1)}$ is *not* identical to the inverse of $G^{(l+1)}$ as can be proven by direct multiplication. Thus updating the inverse Jacobian is indeed different from updating the Jacobian itself, followed by explicit inversion. Actually, this is not quite surprising as the Broyden update is meant as an approximation to the exact matrix (Jacobian or inverse) based on knowledge collected during the iteration process.

According to the original idea of Broyden the updated inverse Jacobian $G^{(l+1)}$ is just the old one ($G^{(l+1-M)}$) with certain matrix elements replaced such that the set of conditions (5.20) is fulfilled. This is also obvious from the last line of Eq. (7.2) where the second term is embraced by projection operators onto the subspaces spanned by the vectors entering Eq. (5.20) and the first term is just the old inverse Jacobian its application being restricted to the subspace orthogonal to the F-space (to which the second term is applied). (In this context the exact form of the initial guess $G^{(l+1-M)}$ does not really matter. It is, however, important that $G^{(l+1-M)}$ contains only knowledge collected prior to the actual update step.) Thus each element of the inverse Jacobian contains either the corresponding ele-

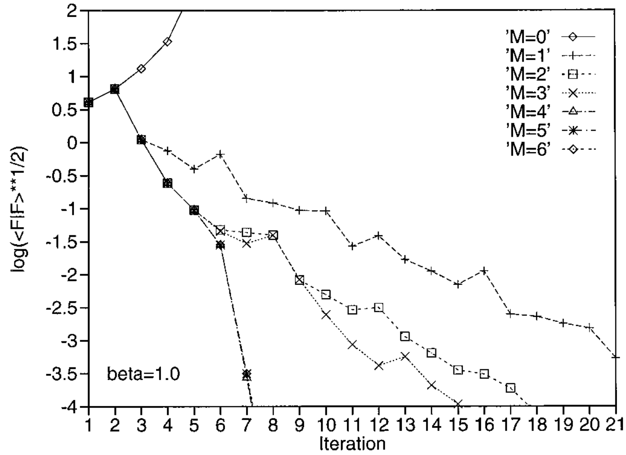

FIG. 5. Rate of convergence using the generalized Broyden method for updating the inverse Jacobian with different numbers M of iterations to be mixed in. The mixing parameter is $\beta = 1.0$.

ment of the old inverse Jacobian $G^{(l+1-M)}$ or else a new one fulfilling Eq. (5.20). Both portions are strictly separated; i.e., there is no element of the inverse Jacobian which contains a mixture of the first and second terms in Eq. (7.2). Of course, the same arguments again hold for the update of the Jacobian itself. However, according to Eq. (5.3) we finally *do* need the inverse Jacobian. Hence in contrast to working directly with the inverse Jacobian we now have to invert the updated Jacobian thus making the fundamental difference between Broyden's first and second method. Since when inverting the updated Jacobian we inevitably remove the just-mentioned separation of the portions corresponding to $J^{(l+1-M)}$ and the conditions (5.19), respectively. This might be related to the fact that the first term in Eq. (7.2) is *not* preceded by a projection operator $(\mathscr{I} - \mathscr{P}_x)$, in which case inversion *would* preserve the separation of the orthogonal subspaces. Hence by inverting the Jacobian we actually end up with a mixture of the two terms contributing to Eq. (7.2) (i.e., its counterpart for $J^{(l+1)}$). As an effect after inversion each element of $(J^{(l+1)})^{-1}$ is a linear combination of terms which could be traced back to either $J^{(l+1-M)}$ or else the matrix arising from the conditions (5.19). Insofar as the situation is not unlike the one described above, where the introduction of weights led to the concurrence of terms which correspond to either simple mixing or full update, in the same manner inversion of the updated Jacobian will also lead to an effective deceleration as compared to updating the inverse Jacobian. Although this will be far from being a drastic effect we still finally note the principal superiority of Broyden's second method over his first one.

In order to get a better feeling for the previous arguments we have implemented the generalized Broyden method as described in Section V for both the inverse Jacobian and the Jacobian itself. The results for different values of M are shown in Figs. 5 and 6 which as expected reveal the small but significant superiority of Broyden's second method. Of course, this slight difference will not have a major influence on applications of these schemes. It is, however, important in the present context of a comparative study of methods.

Next we turn to the comparison of the Anderson mixing and the Broyden update for the inverse Jacobian as formulated in Sections IV and V:

Similar comparisons based on numerical calculations already exist in the literature [17, 18]. However, these authors have only considered the Anderson scheme with at most two previous iterations mixed in; i.e., they have concentrated on the special case $M = 2$. At the same time the modified Broyden method was always used with the full history built in, i.e., $M = l$. Comparisons based on these choices then yielded the not quite surprising result that the modified Broyden scheme is superior to the Anderson mixing.

As concerns theoretical investigations Blügel pointed to interrelations, between the Anderson and the Broyden methods but unfortunately he did not consider these schemes with more than one previous iteration taken into account [22]. This was done by van Leuken who redefined the Anderson scheme in terms of difference vectors $|\Delta x^{(m)}\rangle$ and inserted weights w_m in order to allow for a comparison to the modified Broyden scheme as formulated by Johnson [18]. He then arrived at the same results for both schemes but, due to the weights, he could not observe the full identity and regarded the modified Broyden method as a generalization of the Anderson mixing.

According to van Leuken the whole proof may be reduced to a reformulation of the Anderson mixing in terms of difference vectors, instead of input or residual vectors

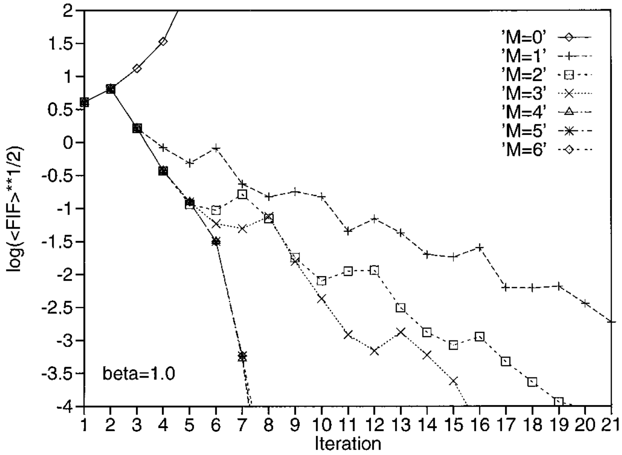

FIG. 6. Same as Fig. 5 but for updating the Jacobian itself followed by inversion.

themselves; thus we start out from Eq. (4.1) and rewrite the general input vector in the following way:

$$\begin{aligned} |\bar{x}^{(l)}\rangle &= |x^{(l)}\rangle + \sum_{j=1}^{M} \vartheta_j^{(l)} \left(|x^{(l-j)}\rangle - |x^{(l)}\rangle\right) \\ &= |x^{(l)}\rangle - \sum_{j=1}^{M} \vartheta_j^{(l)} \sum_{m=l-j}^{l-1} \left(|x^{(m+1)}\rangle - |x^{(m)}\rangle\right) \qquad (7.3) \\ &= |x^{(l)}\rangle - \sum_{m=l-M}^{l-1} \gamma_m^{(l)} |\Delta x^{(m)}\rangle. \end{aligned}$$

Here we have used the definition (5.5) of difference vectors and furthermore abbreviated:

$$\gamma_m^{(l)} = \sum_{j=l-m}^{M} \vartheta_j^{(l)}. \qquad (7.4)$$

The analogous steps performed for the general residual vector yield

$$|\bar{F}^{(l)}\rangle = |F^{(l)}\rangle - \sum_{m=l-M}^{l-1} \gamma_m^{(l)} |\Delta F^{(m)}\rangle. \qquad (7.5)$$

In the same manner as described in Section IV the expansion coefficients entering Eqs. (7.3) and (7.5) are determined by minimizing the norm $E = \langle \bar{F}^{(l)} | \bar{F}^{(l)} \rangle$ of the general residual vector (7.5) with respect to the coefficients $\gamma_m^{(l)}$. This leads to the linear equation system,

$$\sum_{m=l-M}^{l-1} \langle \Delta F^{(n)} | \Delta F^{(m)} \rangle \gamma_m^{(l)} = \langle \Delta F^{(n)} | F^{(l)} \rangle \;\; \forall_{n=l-M,\ldots,l-1} \qquad (7.6)$$

which has to be solved for the coefficients $\gamma_m^{(l)}$. Once these are known the input vector for the following iteration will be constructed according to Eq. (4.4) which in terms of the $\gamma_m^{(l)}$ reads as

$$|x^{(l+1)}\rangle = |x^{(l)}\rangle + \beta^{(l)} |F^{(l)}\rangle - \sum_{m=l-M}^{l-1} \gamma_m^{(l)} [|\Delta x^{(m)}\rangle + \beta^{(l)} |\Delta F^{(m)}\rangle]. \qquad (7.7)$$

With the reformulation of the Anderson mixing at hand we are now able to relate it to the results of the Broyden method which works with difference vectors from the very beginning. To be concrete we compare Eqs. (7.6) and (7.7) to Eqs. (5.31) and (5.30) of the generalized Broyden method thus revealing the full identity of both schemes if the simple mixing ansatz (5.25) is used as an initial guess for the inverse Jacobian of the Broyden method. Yet it should be noted that any other choice would require some information about the system to be iterated and so the simple mixing in any case constitutes the best starting point.

In summary, the Anderson mixing and the generalized second Broyden method as outlined in Section V are fully identical. Taken together with the arguments given in the first part of this section this statement, moreover, holds for the modified second Broyden method when tuned to optimal convergence.

The just-demonstrated identity of the Anderson mixing and the generalized second Broyden method now offer a new point of view for the discussion of weights in the modified second Broyden scheme. As we can tell from Eq. (6.1) these weights could be likewise interpreted as an effective scaling of the difference vectors $|\Delta x^{(m)}\rangle$ and $|\Delta F^{(m)}\rangle$ by a factor $\sqrt{w_m^2/w_0^2 + w_m^2})$. As a consequence, as soon as w_0 is nonzero and all the other weights are not equal, this would result in a change of metric in the spaces spanned by the above vectors. Using now the equivalence to the Anderson mixing which approaches the final solution by minimizing in the residual vector space it becomes obvious that any change in the metric would bring us away from the optimum and thus worsen the effectiveness of the method.

Before we proceed it is worth noting that the above derived formulation of the Anderson mixing is fully equivalent to the one given in Section IV. In particular we point out that the matrices entering Eqs. (4.3) and (7.6) result from each other by applying elementary row and column operations and, hence, have exactly the same determinant. As a consequence whenever there exist linear dependencies in the system of residual vectors which cause the determinant to vanish they would affect both formulations in the same manner and thus lead to the same instabilities. However, we now even realize that this also holds for the Broyden update for which the corresponding linear equation system is given by Eq. (5.31).

Altogether, we realize that regarding any of the two methods as superior is not correct. In contrast, the identity of the formalisms should be related to the completely different starting points the Anderson mixing and the Broyden update take and which offer two different and rather complementary points of view of the same problem. This will be used in the following section and might, in general, be helpful for further developments.

VIII. EXTENDED ANDERSON MIXING

Having witnessed the full identity of the Anderson mixing and the generalized second Broyden update, as well as their superiority over the generalized first Broyden method we are still faced with the situation that linear dependencies in the set of residual vectors might prohibit a proper solution of the linear equation systems (4.3), (5.31), or (7.6), even if this in practice never has been observed by the author. We will now present a quite simple extension of the Anderson mixing which in any case helps to avoid such problems:

Starting from the reformulation of the Anderson mixing presented in Section VII we now prevent the determinant of the overlap matrix in Eq. (7.6) from vanishing by adding a small but finite value to the diagonal elements thus corresponding to w_0^2 in the modified Broyden scheme according to Vanderbilt and Louie. For the Anderson mixing this could easily be achieved by defining a new functional

$$E' = E + w_0^2 \sum_{m=l-M}^{l-1} (\gamma_m^{(l)})^2 a_{mm} \tag{8.1}$$

which when minimized leads to the following linear equation system:

$$\begin{aligned} &\sum_{m=l-M}^{l-1} (1 + w_0^2 \delta_{nm}) \langle \Delta F^{(n)} | \Delta F^{(m)} \rangle \gamma_m^{(l)} \\ &\quad = \langle \Delta F^{(n)} | F^{(l)} \rangle \quad \forall_{n=l-M,\ldots,l-1}. \end{aligned} \tag{8.2}$$

To be pictorial we point out that the uniqueness in the determination of the coefficients $\gamma_m^{(l)}$, if taken away by linear dependencies, can be restored by adding a further “symmetry breaking” condition which, however, must not change the problem we want to solve too drastically. For this reason in practice we put w_0^2 to 10^{-4}.

According to the discussion of the modified Broyden method in the previous section, introducing a small but finite additive diagonal element w_0^2 causes a minor deceleration of the iteration process. Yet above extension of the Anderson mixing does not change the metric in the space spanned by the vectors $|\Delta F^{(m)}\rangle$ and, hence, the linear dependencies have been resolved without affecting the final solution. Furthermore, as test calculations have shown, the results for $w_0^2 = 10^{-4}$ show almost no deviation from those obtained with $w_0^2 = 0$.

In summary, we have finally demonstrated that arguments against using the Anderson mixing with $M > 2$ and which, according to the proof given in the previous section, also apply to the Broyden update for the same M do not really hold since there exist efficient means to avoid such problems.

In order to illustrate the schemes discussed in this work we also applied the extended Anderson mixing to the test case defined in Section II. The results are presented in Figs. 7 and 8 for two different mixing parameters and several values of the number M of previous iterations taken into account. Note, that the curves marked $M = 0$ correspond to simple mixing. Moreover, we point out that the curve for $M = 6$ in Fig. 7 is identical to those obtained for $w_0 = 0.01$ in Figs. 1 and 3, thus reflecting the identity of the Anderson mixing and the generalized second Broyden method shown in Section VII. Due to this identity we furthermore refer to Fig. 5 which would be the same if calculated with the extended Anderson mixing (Actually, Figs. 5 and 6 were calculated with $w_0 = 0.01$ in order to avoid numerical instabilities). Figures 5, 7, and 8 clearly demonstrate a number of appealing features: First, we observe an optimal convergence as represented by the curve for $M = 6$ already for lower values, namely $M = 4$ and $M = 5$ for high and low values of β, respectively. This reflects the opinion expressed by Anderson that the power of the method increases slowly for $M > 3$ since then we would only mix in the poor vectors of early iterations. Following his suggestion to fix M to 4 or 5 we arrive in practice at a good convergence with the vectors of only a few previous iterations to be stored and a linear equation system of low order to be solved.

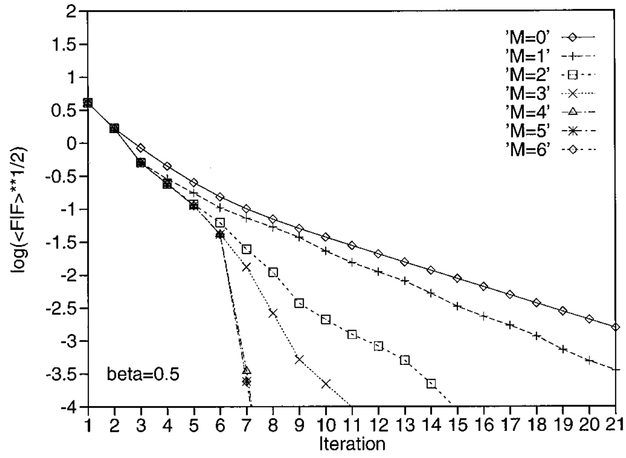

FIG. 7. Rate of convergence using the extended Anderson method with different numbers M of iterations to be mixed in. The mixing parameter is $\beta = 0.5$.

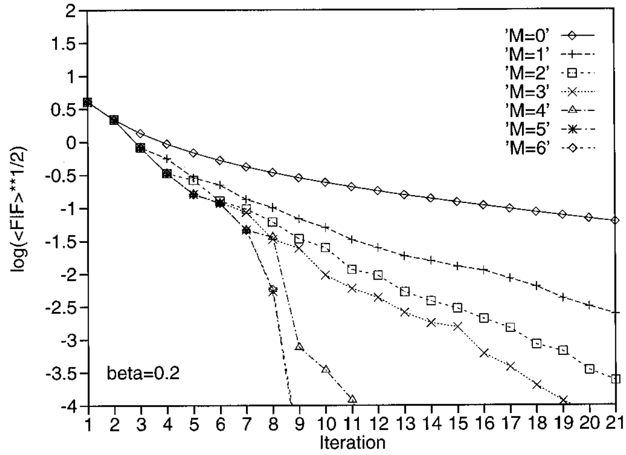

FIG. 8. Same as Fig. 7 but for $\beta = 0.2$.

Second, on comparison of Figs. 5, 7, and 8 we note the high sensitivity on the mixing parameter for the simple mixing and the low M Anderson mixing as well as the decreasing dependence on β for increasing values of M. Thus with the number M being not too small the choice of β becomes rather unimportant. With the identity of the Anderson mixing and the Broyden update at hand all this is in full agreement with the findings of van Leuken.

Finally, the examples also show that β should not be too small. Following these results, as well as the experience of Dederichs and Zeller and van Leuken, we here suggest using $\beta = 0.5$ which has the advantage of damping the oscillations in early iterations. Giving more conclusive recipes for the choice of M and β is beyond the possibilities of our very simple test case, as well as the scope of the present more general paper. Furthermore, our suggestions should in any case be combined with experience with the actual problem at hand.

IX. CONCLUSION

In the present paper we investigated different types of methods aiming at an acceleration of convergence of iterative vector sequences, namely generalized secant schemes and quasi-Newton–Raphson methods. In particular we discussed two prominent and often used members of these classes, the Anderson mixing and the Broyden update, both generalized to the consideration of arbitrarily many previous iterations.

For the generalized Broyden method we presented a new derivation which may be viewed as a straightforward generalization of the original rank-1 update invented by Broyden. Hence, being fully in the spirit of the latter, our new formulation seems to be its most simple generalization.

Furthermore, we investigated in detail the modified Broyden method introduced by Vanderbilt and Louie which also generalizes the rank-1 Broyden update but, in addition, contains weights as a means to increase flexibility. In particular, as we could show if these weights are used to tune the iteration process to its fastest convergence, they in fact cancel out from the formalism and the results of the modified Broyden scheme become identical to those of our much simpler generalized Broyden method.

Comparison of Broyden's first and second methods in their generalized versions clearly revealed the principal superiority of the latter scheme.

Extending the work of van Leuken we could proof the identity of the Anderson mixing and the second Broyden method in their generalized forms. Thus neither of these schemes is superior but in contrast, due to the different underlying concepts, they both offer complementary points of view, which might be helpful for further developments.

Dealing with potential numerical difficulties all these schemes are faced with, we finally proposed an extension to the Anderson mixing which helps to avoid such problems but preserves the good convergence properties of this method.

As concerns the choice of parameters, our test calculations confirmed prior suggestions of Anderson that to use a value of 4 or 5 for the number M of previous iterations taken into account should be optimal. Moreover, for this choice the iteration process becomes rather independent of the mixing parameter β. However, we suggest fixing it to 0.5.

Finally, we point out that all the schemes discussed here not only apply to the solution of nonlinear equation systems but are likewise suited to unconstrained multidimensional minimizations.

APPENDIX: EXPLICIT MATRIX INVERSION

The objective of the appendix is the inversion of the matrix B defined by Eq. (6.5). Dropping some indices this matrix reads as

$$B = w_0^2 \mathscr{I} + \sum_{m=1}^{M} \mathscr{P}_m, \tag{A1}$$

where

$$\mathscr{P}_m := |\tilde{F}_m\rangle\langle\tilde{F}_m| \tag{A2}$$

and

$$|\tilde{F}_m\rangle := \frac{w_m}{\sqrt{a_{mm}}} |F_m\rangle. \tag{A3}$$

As stressed by our notation the operator B is a linear combination of the unity operator and projectors onto the states $|F_m\rangle$ which, however, need not be orthogonal.

Next the general projector onto the whole space spanned by all the $|F_m\rangle$ is defined by

$$\mathscr{P}_F := \sum_{n,k=1}^{M} |\tilde{F}_n\rangle(\tilde{a}^{-1})_{nk}\langle\tilde{F}_k|, \tag{A4}$$

where the matrix $\tilde{a}$ is the overlap matrix (6.10). It can be easily proved that $\mathscr{P}_F$, indeed, has all the properties of a projection operator.

With these preparations at hand we get from Eq. (A1):

$$\begin{aligned} \mathscr{P}_F B \mathscr{P}_F &= w_0^2 \mathscr{P}_F + \sum_{m=1}^{M} \mathscr{P}_m \\ &= \sum_{m,n,k=1}^{M} |\tilde{F}_m\rangle(\tilde{a}^{-1})_{mn}\tilde{b}_{nk}\langle\tilde{F}_k|, \end{aligned} \tag{A5}$$

where the matrix $\tilde{b}$ is defined by Eq. (6.9). Now it can be easily checked that within the space spanned by $|F_m\rangle$ the inverse of this latter matrix is given by

$$\mathscr{P}_F B^{-1} \mathscr{P}_F = \sum_{m,n,k=1}^{M} |\tilde{F}_m\rangle (\tilde{a}^{-1})_{mn} (\tilde{b}^{-1})_{nk} \langle \tilde{F}_k|. \tag{A6}$$

With these identities at hand it is straightforward now to invert the operator B; using Eq. (6.7) we have

$$\begin{aligned} B^{-1} &= \frac{1}{w_0^2}[\mathscr{I} - \mathscr{P}_F + w_0^2 \mathscr{P}_F B^{-1} \mathscr{P}_F] \\ &= \frac{1}{w_0^2}[\mathscr{I} - \mathscr{P}_F (B - w_0^2 \mathscr{I}) B^{-1} \mathscr{P}_F] \\ &= \frac{1}{w_0^2}\left[\mathscr{I} - \sum_{m=1}^{M} \mathscr{P}_m \mathscr{P}_F B^{-1} \mathscr{P}_F\right] \\ &= \frac{1}{w_0^2}\left[\mathscr{I} - \sum_{m,k=1}^{M} |\tilde{F}_m\rangle (\tilde{b}^{-1})_{mk} \langle \tilde{F}_k|\right]. \end{aligned} \tag{A7}$$

REFERENCES

1. J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables* (Academic Press, New York/London, 1970).
2. E. Polak, *Computational Methods in Optimization* (Academic Press, New York, 1971).
3. G. D. Byrne and C. A. Hall, (Eds.), *Numerical Solution of Systems of Nonlinear Algebraic Equations* (Academic Press, New York/London, 1973).
4. D. A. H. Jacobs, (Ed.), *The State of the Art in Numerical Analysis* (Academic Press, London, 1977).
5. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes—The Art of Scientific Computing* (Cambridge Univ. Press, Cambridge, 1989).
6. J. Stoer, *Numerische Mathematik* (Springer-Verlag, Berlin, 1993).
7. P. Hohenberg and W. Kohn, *Phys. Rev. B* **136,** 864 (1964).
8. W. Kohn and L. J. Sham, *Phys. Rev. A* **140,** 1133 (1965).
9. J. Kübler and V. Eyert, ''Electronic Structure Calculations,'' in *Electronic and Magnetic Properties of Metals and Ceramics*, edited by K. H. J. Buschow, (*VCH* Verlagsgesellschaft, Weinheim, 1991), p. 1; *Materials Science and Technology*, Vol. 3A edited by R. W. Cahn, P. Haasen, E. J. Kramer (VCH Verlagsgesellschaft, Weinheim).
10. R. Zeller, *Modelling Simul. Mater. Sci. Eng.* **1,** 553 (1993).
11. C. G. Broyden, *Math. Comput.* **19,** 577 (1965).
12. J. E. Dennis, Jr. and J. J. Moré, *SIAM Rev.* **19,** 46 (1977).
13. P. Bendt and A. Zunger, *Phys. Rev. B* **26,** 3114 (1982).
14. G. P. Srivastava, *J. Phys. A* **17,** L317 (1984).
15. D. Singh, H. Krakauer, and C. S. Wang, *Phys. Rev. B* **34,** 8391 (1986).
16. D. Vanderbilt and S. G. Louie, *Phys. Rev. B* **30,** 6118 (1984).
17. D. D. Johnson, *Phys. Rev. B* **38,** 12807 (1988).
18. H. van Leuken, Ph.D. thesis, University of Amsterdam, 1991 (unpublished).
19. D. G. Anderson, *J. Assoc. Comput. Mach.* **12,** 547 (1965).
20. P. H. Dederichs and R. Zeller, *Phys. Rev. B* **28,** 5462 (1983).
21. L. F. Mattheiss and D. R. Hamann, *Phys. Rev. B* **33,** 823 (1986).
22. S. Blügel, Ph.D. thesis, Rheinisch-Westfälische Technische Hochschule Aachen, 1987; Berichte der Kernforschungsanlage Jülich Nr. 2197.