JSIAM Letters Vol.1 (2009) pp.32–35 ©2009 Japan Society for Industrial and Applied Mathematics

Reformulation of the Anderson method using singular value decomposition for stable convergence in self-consistent calculations

Akitaka Sawamura[1]

Sumitomo Electric Industries, Ltd., 1-1-3, Shimaya, Konohana-ku, Osaka 554-0024, Japan[1]

E-mail *sawamura-akitaka@sei.co.jp*

Received March 18, 2009, Accepted April 20, 2009

Abstract

The Anderson method provides a significant acceleration of convergence in solving nonlinear simultaneous equations by trying to minimize the residual norm in a least-square sense at each iteration step. In the present study I use singular value decomposition to reformulate the Anderson method. The proposed version contains only a single parameter which should be determined in a trial-and-error way, whereas the original one contains two. This reduction leads to stable convergence in real-world self-consistent electronic structure calculations.

Keywords nonlinear simultaneous equations, least-square method, the Broyden method, the Pulay method, electronic-structure calculations

Research Activity Group Algorithms for Matrix / Eigenvalue Problems and their Applications

1. Introduction

In the past few years first-principles calculations based on density functional theory [1] have gained enormous interest among solid-state physists, materials scientists, and quantum chemists. The Kohn-Sham equation [2], which plays a vital role within the density functional theory, is not only an eigenvalue problem, but also an implicitly defined, nonlinear fixed-point problem of interelecton potential [3–5] at least when local density approximation [2] is introduced. In other words, the Kohn-Sham equation is solved when the self-consistent interelecton potential is found. The Anderson method [6] is frequently employed for this purpose. It should be noted that the Pulay method [7] and limited-memory modifications [8–11] of the second Broyden method [12] are essentially equivalent to the Anderson method [13], while the first Broyden method can also be cast into the limited-memory form [14, 15].

Suppose that for a system of nonlinear equations $\vec{\mathfrak{F}}(\vec{x}) = \vec{0}$, there are independent variable column vectors,

$$\{\vec{x}_n, \vec{x}_{n-1}, \ldots, \vec{x}_{n-k}\},$$

which are hopefully approaching a solution, and accompanying residual column vectors,

$$\{\vec{y}_n, \vec{y}_{n-1}, \ldots, \vec{y}_{n-k}\},$$

where subscripts denote iteration steps. In a simple iteration method, the independent vector at the $(n+1)$th iteration step is given by

$$\vec{x}_{n+1} = \vec{x}_n + \alpha \vec{y}_n, \tag{1}$$

where α is a mixing factor ranging from a scalar to a preconditioning matrix [16, 17] to a nonlinear procedure [18,19]. In the Anderson method, however, a virtual residual vector,

$$\vec{y}_n^\star = \vec{y}_n + \sum_{1\le\nu\le k} \gamma_\nu \frac{\vec{y}_{n-\nu+1} - \vec{y}_{n-\nu}}{\|\vec{y}_{n-\nu+1} - \vec{y}_{n-\nu}\|}, \tag{2}$$

is introduced. Here γ_ν are parameters to be so determined that the virtual residual norm $\|\vec{y}_n^\star\|$ is minimized in a least-square sense. Then an accompanying virtual independent vector,

$$\vec{x}_n^\star = \vec{x}_n + \sum_{1\le\nu\le k} \gamma_\nu \frac{\vec{x}_{n-\nu+1} - \vec{x}_{n-\nu}}{\|\vec{y}_{n-\nu+1} - \vec{y}_{n-\nu}\|}, \tag{3}$$

is defined on assumption of linearity. $\vec{x}_n^\star$ is expected to be a minimizer for $\left\|\vec{\mathfrak{F}}\right\|$ within the available subspace $\{\vec{x}_n, \vec{x}_{n-1}, \ldots, \vec{x}_{n-k}\}$. Last, the independent variable vector for the next step is predicted by applying the simple iteration method to $\vec{x}_n^\star$ and $\vec{y}_n^\star$ as

$$\vec{x}_{n+1} = \vec{x}_n^\star + \alpha \vec{y}_n^\star. \tag{4}$$

In practice, a specialized linear solver should be used to determine the parameters γ_ν reliably without encountering numerical instability. This means that a maximum condition number must be set for the linear solver beforehand. Moreover a limit for the number of the previous independent and residual vectors considered must be also set beforehand. Since the two parameters cannot be obtained a priori, they are determined in an ad hoc way. In the present study I eliminate the latter by reformulating the Anderson method based on singular value decomposition (SVD) [20]. This makes application of the Anderson method a little easier. Furthermore, stable convergence is achieved in the sense that the numbers of iteration steps required by self-consistency are

less sensitive with respect to the remaining parameters as confirmed by test calculations.

2. Conventional Method

For simplicity I define a rectangular matrix as

$$Y_n = \left\{ \frac{\vec{y}_n - \vec{y}_{n-1}}{\|\vec{y}_n - \vec{y}_{n-1}\|}, \frac{\vec{y}_{n-1} - \vec{y}_{n-2}}{\|\vec{y}_{n-1} - \vec{y}_{n-2}\|}, \ldots, \frac{\vec{y}_{n-k+1} - \vec{y}_{n-k}}{\|\vec{y}_{n-k+1} - \vec{y}_{n-k}\|} \right\}, \tag{5}$$

and a column vector containing γ_ν as

$$\Gamma = \left\{ \begin{array}{c} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_k \end{array} \right\}. \tag{6}$$

I omit the right-pointing arrow above Γ to emphasize that in general Γ is different form x_ν and y_ν in the number of rows. Using Y_n and Γ, (2) is rewritten as

$$\vec{y}_n^\star = \vec{y}_n + Y_n \Gamma. \tag{7}$$

The formal solution of Γ which minimizes $\|\vec{y}_n^\star\|$ is given by

$$\Gamma = -\left(Y_n^T Y_n\right)^{-1} Y_n^T \vec{y}_n. \tag{8}$$

Determining Γ using (8) literally should be discouraged, because of the potentially large condition number of $Y_n^T Y_n$. Instead, Γ is computed in a following way. First, if the SVD is employed, Y_n is factorized into

$$Y_n = \mathcal{Y}_n \Sigma_n V_n^T, \tag{9}$$

where $\mathcal{Y}_n$ and V_n are matrices containing the left and right singular vectors of Y_n, respectively, while Σ_n is a diagonal matrix of the singular values. Then a corresponding truncated factorization,

$$Y_n \approx \overline{Y}_n = \overline{\mathcal{Y}}_n \overline{\Sigma}_n \overline{V}_n^T, \tag{10}$$

is considered. Here $\overline{\Sigma}_n$ is a diagonal matrix of the l largest singular values of Σ_n, while $\overline{\mathcal{Y}}_n$ and $\overline{V}_n^T$ contain the l column vectors of $\mathcal{Y}_n$ and the l row vectors of V_n^T corresponding to the l largest singular values, respectively. l, the effective rank of Y_n, is the largest integer so determined that the condition number of $\overline{\Sigma}_n$ does not exceed the first predetermined limit $s_{\max}$. Of course, l can be equal to k. Last, Γ is given by

$$\Gamma = -\overline{V}_n \overline{\Sigma}_n^{-1} \overline{\mathcal{Y}}_n^T \vec{y}_n. \tag{11}$$

At the next iteration step, Y_{n+1} may be set to be

$$\begin{aligned} Y_{n+1} &= \left\{ \frac{\vec{y}_{n+1} - \vec{y}_n}{\|\vec{y}_{n+1} - \vec{y}_n\|}, Y_n \right\} \\ &= \left\{ \frac{\vec{y}_{n+1} - \vec{y}_n}{\|\vec{y}_{n+1} - \vec{y}_n\|}, \frac{\vec{y}_n - \vec{y}_{n-1}}{\|\vec{y}_n - \vec{y}_{n-1}\|}, \ldots, \frac{\vec{y}_{n-k+1} - \vec{y}_{n-k}}{\|\vec{y}_{n-k+1} - \vec{y}_{n-k}\|} \right\}. \end{aligned} \tag{12}$$

The usual practice is, however, that if k has reached the second predetermined limit $k_{\max}$, the rightmost (oldest) column of the right-hand side of (12) is removed as

$$Y_{n+1} = \left\{ \frac{\vec{y}_{n+1} - \vec{y}_n}{\|\vec{y}_{n+1} - \vec{y}_n\|}, \frac{\vec{y}_n - \vec{y}_{n-1}}{\|\vec{y}_n - \vec{y}_{n-1}\|}, \ldots, \frac{\vec{y}_{n-k+2} - \vec{y}_{n-k+1}}{\|\vec{y}_{n-k+2} - \vec{y}_{n-k+1}\|} \right\}, \tag{13}$$

to avoid excessive growth.

3. Proposed Method

Along with (5), I define a rectangular matrix containing the independent variable vectors as

$$X_n = \left\{ \frac{\vec{x}_n - \vec{x}_{n-1}}{\|\vec{y}_n - \vec{y}_{n-1}\|}, \frac{\vec{x}_{n-1} - \vec{x}_{n-2}}{\|\vec{y}_{n-1} - \vec{y}_{n-2}\|}, \ldots, \frac{\vec{x}_{n-k+1} - \vec{x}_{n-k}}{\|\vec{y}_{n-k+1} - \vec{y}_{n-k}\|} \right\}. \tag{14}$$

Since $\overline{\mathcal{Y}}_n = Y_n \overline{V}_n \overline{\Sigma}_n^{-1}$ holds, a similar quantity,

$$\overline{\mathcal{X}}_n = X_n \overline{V}_n \overline{\Sigma}_n^{-1}, \tag{15}$$

is introduced. $\vec{x}_n^\star$ and $\vec{y}_n^\star$ are computed by working with $\overline{\mathcal{X}}_n$ and $\overline{\mathcal{Y}}_n$ as

$$\vec{x}_n^\star = \vec{x}_n + \overline{\mathcal{X}}_n \Gamma' \tag{16}$$

and

$$\vec{y}_n^\star = \vec{y}_n + \overline{\mathcal{Y}}_n \Gamma', \tag{17}$$

respectively, where Γ' is obtained by

$$\Gamma' = -\overline{\mathcal{Y}}_n^T \vec{y}_n. \tag{18}$$

At the next iteration step, X_{n+1} and Y_{n+1} are updated by

$$X_{n+1} = \left\{ \frac{\vec{x}_{n+1} - \vec{x}_n}{\|\vec{y}_{n+1} - \vec{y}_n\|}, \overline{\mathcal{X}}_n \right\} \tag{19}$$

and

$$Y_{n+1} = \left\{ \frac{\vec{y}_{n+1} - \vec{y}_n}{\|\vec{y}_{n+1} - \vec{y}_n\|}, \overline{\mathcal{Y}}_n \right\}, \tag{20}$$

respectively. $\overline{\mathcal{X}}_n$ and $\overline{\mathcal{Y}}_n$ consist of the l column vectors while by construction $l \leq k$ holds. Therefore X_{n+1} and Y_{n+1} are unlikely to fatten endlessly even if no limit is imposed. No column in X_{n+1} and Y_{n+1} has to be discarded artificially.

Since $\overline{\mathcal{Y}}_n$ represents a numerically effective subspace spanned by Y_n, replacing (12) with (20) makes little information carried in Y_{n+1} be lost even in the case of $l < k$. This no longer holds when the leftmost column is discarded as in the case of (13). Nevertheless if $k_{\max}$ is set too large with $s_{\max}$ kept moderate in the conventional method, the predicted x_{n+1} may be contaminated by excessively old x_ν and y_ν, because Γ in (11) is a *minimum-norm* least-square solution. Therefore the proposed method is expected to outperform the conventional one.

4. Test Calculations

The conventional and proposed methods are compared by applying them to first-principles calculations for wurzite ZnO based on plane-wave, pseudopotential

Table 1. Iterations required to reach self-consistency for wurzite ZnO with various maximum singular values $s_{\max}$ and history data limits $k_{\max}$. Maximum numbers of history data reached in the proposed method are shown in parentheses. Note that lattice parameters and atomic positions in the unit cell are optimized.

$\alpha = 0.2$

$1/s_{\max}$	$k_{\max}$	3×10^{-1}	1×10^{-1}	3×10^{-2}	1×10^{-2}	3×10^{-3}	1×10^{-3}	3×10^{-4}	1×10^{-4}
Conventional	5	74	67	65	49	51	48	50	50
	10	74	54	56	46	47	42	42	45
	20	55	62	51	47	47	42	47	43
	40	56	62	54	65	56	48	46	44
Proposed		43(14)	42(14)	42(15)	45(20)	52(27)	42(28)	42(28)	42(28)

$\alpha = 0.4$

$1/s_{\max}$	$k_{\max}$	3×10^{-1}	1×10^{-1}	3×10^{-2}	1×10^{-2}	3×10^{-3}	1×10^{-3}	3×10^{-4}	1×10^{-4}
Conventional	5	53	52	46	45	45	45	44	44
	10	50	45	44	42	39	38	37	38
	20	58	42	44	47	42	41	40	39
	40	46	58	46	40	43	39	41	41
Proposed		38(12)	40(13)	36(15)	40(21)	40(23)	42(26)	41(27)	41(27)

$\alpha = 0.8$

$1/s_{\max}$	$k_{\max}$	3×10^{-1}	1×10^{-1}	3×10^{-2}	1×10^{-2}	3×10^{-3}	1×10^{-3}	3×10^{-4}	1×10^{-4}
Conventional	5	39	33	33	31	32	31	32	32
	10	36	34	34	31	30	31	31	31
	20	35	35	37	35	33	32	34	34
	40	37	35	35	34	32	33	34	34
Proposed		32(6)	31(11)	32(16)	33(16)	34(20)	34(20)	34(20)	34(20)

$\alpha = 1.6$

$1/s_{\max}$	$k_{\max}$	3×10^{-1}	1×10^{-1}	3×10^{-2}	1×10^{-2}	3×10^{-3}	1×10^{-3}	3×10^{-4}	1×10^{-4}
Conventional	5	100	59	50	46	45	45	45	45
	10	139	59	42	40	40	37	39	39
	20	167	63	45	40	41	37	38	37
	40	72	68	71	47	45	37	38	40
Proposed		41(13)	37(13)	38(18)	38(20)	39(23)	40(25)	40(26)	40(26)

approach [21, 22]. Lattice parameters and atomic positions in the unit cell are also optimized. Remaining technical details are explained elsewhere [23]. The mixing factor α is chosen to be a scalar parameter.

The parameters and results are shown in Table 1. For $\alpha = 0.8$, both the methods have achieved fast convergence of iteration steps below 40. Clearly, however, the proposed method is the less sensitive to the selection of α and $s_{\max}$. Almost always the self-consistency is reached within about 40 steps. In contrast, when the parameters are chosen poorly, for example at $\alpha = 1.6$ and $1/s_{\max} = 3 \times 10^{-1}$, the conventional method requires more than 100 iteration steps depending on $k_{\max}$. More importantly, finding the optimal $k_{\max}$ seems to be difficult, because though the iteration steps increase with $k_{\max} \leq 20$, the fastest convergence is achieved at $k_{\max} = 40$. As a whole whereas the larger $s_{\max}$ is desirable for the conventional method, a guiding principle for $k_{\max}$ is unclear. Table 1 shows also the maximum l reached within the proposed method. These values might be taken as the best $k_{\max}$ for the conventional method. At $k_{\max}$ near these values, however, the conventional method does not necessarily show the comparable performance of the proposed one. This is likely because discarding the oldest column is not the best strategy to keep X_n and Y_n from excessive growth as pointed out in the previous section.

5. Conclusion

Reformulation of the Anderson method for a system of nonlinear equations has been described. The Anderson method in practice requires two empirical parameters commanding to what extent stably the least-square problem appearing at each iteration step is solved and how many vectors containing the convergence history information are retained. In the proposed method the SVD is used to extract the effective information from the history vectors, rather than as a black-box tool for solving the least-square problem. The extracted vectors are chosen to play a role of storage space for the history information. Thereby the latter empirical parameter is no more needed. This makes the proposed method be the less sensitive to the selection of the remaining parameter and the mixing factor and the more efficient because of a smarter way of discarding a redundant part of the history information, as supported by the stable convergence in the test calculations.

References

[1] P. Hohenberg and W. Kohn, Inhomogeneous electron gas, Phys. Rev. **136** (1964), B864–B871.

[2] W. Kohn and L. J. Sham, Self-consistent equations including exchange and correlation effects, Phys. Rev., **140** (1965), A1133–A1138.

[3] P. Bendt and A. Zunger, New approach for solving the density-functional self-consistent-field problem, Phys. Rev. B, **26** (1982), 3114–3137.

[4] J. F. Annett, Efficiency of algorithms for Kohn-Sham density functional theory, Comput. Mater. Sci., **4** (1995), 23–42.

[5] X. Gonze, Toward a potential-based conjugate gradient algorithm for order-N self-consistent total energy calculations, Phys. Rev. B, **54** (1996), 4383–4386.

[6] D. G. Anderson, Iterative procedures for nonlinear integral equations, J. Assoc. Comput. Mach., **12** (1965), 547–560.

[7] P. Pulay, Convergence acceleration of iterative sequence: the case of SCF iteration, Chem. Phys. Lett., **73** (1980), 393–398.

[8] G. P. Srivastava, Broyden's method for self-consistent field convergence acceleration, J. Phys. A: Math. Gen., **17** (1984), L317–L321.

[9] D. Vanderbilt and S. G. Louie, Total energies of diamond (111) surface reconstructions by a linear combination of atomic orbitals method, Phys. Rev. B, **30** (1984), 6118–6130.

[10] D. D. Johnson, Modified Broyden's method for accelerating convergence in self-consistent calculations, Phys. Rev. B, **38** (1988), 12807–12813.

[11] A. Sawamura, M. Kohyama, T. Keishi, M. Kaji, Acceleration of self-consistent electronic-structure calculations: Storage-saving and multiple-secant implementation of the Broyden method, Mater. Trans., JIM, **40** (1999), 1186–1192.

[12] C. G. Broyden, A class of methods for solving nonlinear simultaneous equations, Math. Comput., **19** (1965), 577–593.

[13] V. Eyert, A comparative study on methods for convergence acceleration of iterative vector sequences, J. Comput. Phys., **124** (1996), 271–285.

[14] M. S. Engelman, G. Strang and K. -J. Bathe, The application of quasi-Newton method in fluid mechanics, Intern. J. Numer. Meth. Eng., **17** (1981), 707–718.

[15] R. H. Byrd, J. Nocedal and R. B. Schnabel, Representations of quasi-Newton matrices and their use in limited memory methods, Math. Progm., **63** (1994), 129–156.

[16] G. P. Kerker, Efficient iteration scheme for self-consistent pseudopotential calculations, Phys. Rev. B, **23** (1981), 3082–3084.

[17] K. -M. Ho, J. Ihm and J. D. Joannopoulos, Dielectric matrix scheme for fast convergence in self-consistent electronic-structure calculations, Phys. Rev. B, **25** (1982), 4260–4262.

[18] D. Raczkowski, A. Canning and L. W. Wang, Thomas-Fermi charge mixing for obtaining self-consistency in density functional calculations, Phys. Rev. B, **64** (2001), 121101-1–121101-4.

[19] A. Sawamura and M. Kohyama, A second-variational prediction operator for fast convergence in self-consistent electronic structure calculations, Mater. Trans., JIM, **45** (2004), 1422–1428.

[20] G. H. Golub and C. F. van Loan, Matrix computations, Second edition, John Hopkins Univ. Press, London, 1989.

[21] J. Ihm, A. Zunger and M. L. Cohen, Momentum-space formalism for the total energy of solids, J. Phys. C: Solid State Phys., **12** (1979), 4409–4422.

[22] W. E. Pickett, Pseudopotential methods in condensed matter applications, Comput. Phys. Rep., **9** (1989), 115–197.

[23] A. Sawamura, M. Kohyama and T. Keishi, An efficient preconditioning scheme for plane-wave-based electronic structure calculations, Comput. Mater. Sci., **14** (1999), 4–7.