Theoret. Chim. Acta (Berl.) 33, 201—214 (1974)

The Convergence Properties of Direct Methods of Energy Minimization with Respect to Linear Coefficients in the LCAO-MO-SCF Approach

B. T. Sutcliffe*

Kemisk Laboratorium IV, H. C. Ørsted Institutet, 2100 Copenhagen Ø, Denmark

Received August 3, 1973

It is shown that in the LCAO-MO-SCF problem, if the molecular orbital orthonormality constraints are introduced in the manner first suggested by Fletcher, then the Hessian of the problem is singular. It is suggested that this singularity may well account for the slow convergence observed using direct energy minimization methods to solve the SCF problem. Ways of avoiding the consequences of this singularity are discussed.

Key words: Direct energy minimization – Convergence of LCAO-MO-SCF calculations

1. Introduction

Recently interest has been revived in "direct" methods of minimizing the energy with respect to such parameters as nuclear position, orbital exponents and orbital (linear) coefficients, following the pioneering work of McWeeny [1, 2] using the steepest descent methods. In particular Fletcher [3] showed how it was possible to use one of the more modern conjugate-direction techniques in such direct minimization, and an approach similar to that of Fletcher was later exploited by Kari and Sutcliffe [4] and by Claxton and Smith [5]. Claxton and Smith concentrated on optimizing the linear coefficients in an unrestricted Hartree-Fock (UHF) approach for systems which had proved convergent only with difficulty using more conventional techniques. Though Claxton and Smith were able to obtain convergence using a direct method (in fact the Fletcher-Reeves method [6]), they commented that the method proceeded only very slowly and could not compete with conventional methods when these methods worked. The object of this paper is to try explain why it is that direct methods have so far proved so disappointing for linear coefficients in closed and in unrestricted LCAO-MO-SCF calculations.

2. Direct Methods of Minimization

Most modern direct minimization techniques are based on choosing a sequence of directions in the co-ordinate space in which the function $f(x)$ is to be minimized, and finding a sequence of points by minimizing, or at least de-

* Permanent address: Department of Chemistry, University of York, York Y01 5DD, England.

creasing, the function value along the chosen direction until a minimum point is found. The most effective of the modern methods are based on the supposition that sufficiently close to the minimum the objective function $f(\boldsymbol{x})$ can be expanded in a Taylor series to second order.

$$f(\boldsymbol{x}) = a + \boldsymbol{b}^T \boldsymbol{x} + \tfrac{1}{2} \boldsymbol{x}^T \boldsymbol{H} \boldsymbol{x} \tag{2.1}$$

where $\boldsymbol{x}$ is a column vector of co-ordinates and the matrix $\boldsymbol{H}$ (the Hessian matrix at the minimum) is assumed to be a real symmetric, positive-definite, non-singular matrix. If this is possible the gradient of the function $g(\boldsymbol{x})$ may be written

$$\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{b} + \boldsymbol{H} \boldsymbol{x} . \tag{2.2}$$

with $g_i(x) = \partial f / \partial x_i$, and hence the minimum point $\boldsymbol{a}^0$ may be found from any arbitrary point a as

$$(\boldsymbol{a}^0 - \boldsymbol{a}) = -\boldsymbol{H}^{-1} \boldsymbol{g}(\boldsymbol{a}) \tag{2.3}$$

provided that $\boldsymbol{H}$ is non-singular.

It can be shown that this problem can be solved in just n steps without having to invert $\boldsymbol{H}$ or indeed without explicit knowledge even of $\boldsymbol{H}$, by constructing a sequence of conjugate-directions, that is a sequence of directions $\boldsymbol{p}_i$ such that

$$\boldsymbol{p}_i^T \boldsymbol{H} \boldsymbol{p}_j = 0 , \quad i \neq j , \tag{2.4}$$

$$\boldsymbol{g}_{i+1}^T \boldsymbol{p}_i = 0 , \tag{2.5}$$

and minimizing the function along these directions. That is, if at any point $\boldsymbol{x} = \boldsymbol{a}$, one knows the direction $\boldsymbol{p} \equiv \boldsymbol{p}(\boldsymbol{a})$, then one constructs the function

$$F(\lambda) = f(\boldsymbol{a} + \lambda \boldsymbol{p}) \tag{2.6}$$

and finds the value of λ, α say, that minimizes $F(\lambda)$ and then the next point in the descent sequence, $\hat{\boldsymbol{a}}$, is chosen according to

$$\hat{\boldsymbol{a}} = \boldsymbol{a} + \alpha \boldsymbol{p} . \tag{2.7}$$

It is easy to see that at the point $\boldsymbol{a}$

$$\hat{\boldsymbol{g}}^T \boldsymbol{p} = 0 \tag{2.8}$$

where

$$\hat{\boldsymbol{g}} \equiv \boldsymbol{g}(\hat{\boldsymbol{a}}) \tag{2.9}$$

and that therefore

$$f(\hat{\boldsymbol{a}}) - f(\boldsymbol{a}) = -(\boldsymbol{g}^T \boldsymbol{p})^2 / 2 \boldsymbol{p}^T \boldsymbol{H} \boldsymbol{p} \tag{2.10}$$

and that

$$\alpha = -\boldsymbol{g}^T \boldsymbol{p} / \boldsymbol{p}^T \boldsymbol{H} \boldsymbol{p} . \tag{2.11}$$

Many methods are available for choosing such conjugate direction, examples are the method of Powell [7] which does not use the gradient matrix, the method of Fletcher and Reeves [6] which uses the gradient matrix and Fletcher and Powell's modification of Davidon's method [8], which uses the gradient matrix

and also yields an estimate of the inverse Hessian at the minimum. A general discussion of such methods in the case of quadratic functions may be found in Huang [9] see also Dixon [10].

It is clear however from the above discussion that considerable difficulties may arise in utilizing one of these methods if the Hessian matrix is not positive definite. Thus in this case it is possible that the denominators in (2.10) and (2.11) become zero so that the location of a minimum in the direction $\boldsymbol{p}$ is just not possible. Even if the methods do not fail overtly because of this, it is the case that proof of quadratic termination for the methods depends on the positive definiteness of $\boldsymbol{H}$ so that one might well expect poor convergence even when the method does not fail outright.

We shall now show that the Hessian at the minimum in the closed shell SCF problem is indeed singular and we suggest that this may be the origin of the difficulties experienced by Claxton and Smith and others. The demonstration we use may be generalized immediately to the UHF problem.

3. The LCAO-MO Closed Shell Problem

Using the notation of McWeeny and Sutcliffe [11] the energy function in the closed shell problem may be written as

$$E = 2\,\mathrm{tr}\,\boldsymbol{h}\boldsymbol{R} + \mathrm{tr}\,\boldsymbol{G}(\boldsymbol{R})\,\boldsymbol{R} \tag{3.1}$$

where $\boldsymbol{h}$ is the matrix of one electron integrals and $\boldsymbol{G}(\boldsymbol{R})$ the usual electron interaction matrix, both in the atomic orbital basis. The matrix $\boldsymbol{R}$ is defined as

$$\boldsymbol{R} = \boldsymbol{T}\boldsymbol{T}^T \tag{3.2}$$

where $\boldsymbol{T}$ is the m by n matrix that relates the n doubly-occupied molecular orbitals to the chosen atomic orbital basis (η).

This function as it stands is not a suitable object for use in a direct minimisation procedure since the variables of the problem, the linear coefficients, T_{ir}, are constrained by the orthonormality requirements among the molecular orbitals, namely

$$\boldsymbol{T}^T\boldsymbol{S}\boldsymbol{T} = \boldsymbol{I} \tag{3.3}$$

where $\boldsymbol{S}$ is the overlap matrix in the atomic orbital basis and $\boldsymbol{I}$ is the n dimensional unit matrix.

These constraints can be incorporated, as was first shown by Fletcher [3], by writing

$$\boldsymbol{T} = \boldsymbol{Y}\boldsymbol{U} \tag{3.4}$$

where $\boldsymbol{Y}$ is an m by n matrix of unconstrained variables and the n by n matrix $\boldsymbol{U}$ is chosen to supply the constraints. In terms of (3.3) it is seen that $\boldsymbol{U}$ must satisfy the equation

$$\boldsymbol{U}\boldsymbol{U}^T = (\boldsymbol{Y}^T\boldsymbol{S}\boldsymbol{Y})^{-1} \tag{3.5}$$

and for the sake of brevity we denote $(\boldsymbol{Y}^T\boldsymbol{S}\boldsymbol{Y})$ by $\boldsymbol{A}$. It follows therefore that we may write

$$\boldsymbol{R}=\boldsymbol{Y}\boldsymbol{A}^{-1}\boldsymbol{Y}^T \tag{3.6}$$

and since the energy depends only on $\boldsymbol{R}$, we see that it is unnecessary to specify $\boldsymbol{U}$ more closely than by (3.5).

Following Fletcher [3] (see also Kari and Sutcliffe [12]) we may determine the gradient of E with respect to the variables Y_{ir} by noting that under the change $\boldsymbol{Y}\rightarrow\boldsymbol{Y}+\delta\boldsymbol{Y}$ such that $E\rightarrow E+\delta E$, then

$$\boldsymbol{R}\rightarrow\boldsymbol{R}+\boldsymbol{Y}\boldsymbol{A}^{-1}\delta\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{S}\boldsymbol{R})+(\boldsymbol{I}-\boldsymbol{R}\boldsymbol{S})\,\delta\boldsymbol{A}^{-1}\boldsymbol{Y}^T\,, \tag{3.7}$$

$$\boldsymbol{A}^{-1}\rightarrow\boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\delta\boldsymbol{A}\boldsymbol{A}^{-1}\,, \tag{3.8}$$

where

$$\delta\boldsymbol{A}=\delta\boldsymbol{Y}^T\boldsymbol{S}\boldsymbol{Y}+\boldsymbol{Y}^T\boldsymbol{S}\delta\boldsymbol{Y} \tag{3.9}$$

and that

$$\boldsymbol{G}(\boldsymbol{R})\rightarrow\boldsymbol{G}(\boldsymbol{R})+\boldsymbol{G}(\delta\boldsymbol{R})\,. \tag{3.10}$$

After a little manipulation it may be shown that for real $\boldsymbol{Y}$, that

$$\delta E=4\,\mathrm{tr}(\boldsymbol{I}-\boldsymbol{S}\boldsymbol{R})\boldsymbol{f}\boldsymbol{Y}\boldsymbol{A}^{-1}\delta\boldsymbol{Y}^T \tag{3.11}$$

where

$$\boldsymbol{f}=\boldsymbol{h}+\boldsymbol{G}(\boldsymbol{R})\,, \tag{3.12}$$

and hence

$$\frac{\partial E}{\partial Y_{ir}}=4\{(\boldsymbol{I}-\boldsymbol{S}\boldsymbol{R})\boldsymbol{f}\boldsymbol{Y}\boldsymbol{A}^{-1}\}_{ir} \tag{3.13}$$

so that the gradient can in this case be represented by an m by n matrix

$$\boldsymbol{V}=4(\boldsymbol{I}-\boldsymbol{S}\boldsymbol{R})\boldsymbol{f}\boldsymbol{Y}\boldsymbol{A}^{-1} \tag{3.14}$$

with the row indices labelling the atomic, and the column indices the molecular orbitals.

It would thus seem that choice of the elements of $\boldsymbol{Y}$ as the variables of minimization according to (3.4) is an extremely good choice since one is able to express the gradient of the energy in a compact manner in terms of them. Furthermore, as they are peculiarly suitable for a direct minimization procedure precisely because they are unconstrained variables and so do not need to be modified to satisfy an ancillary condition at each iteration. As was pointed out by Fletcher [3] if one chooses a variable set subject to an ancillary condition which one needs to restore at the end of an iteration (for example if one chose $\boldsymbol{T}$) and restored orthonormality one cannot use a direct minimization process because the information from the previous iteration is "spoiled" by the restoration of constraints and so the advantages of direct minimization are lost. However the advantages of many direct minimization procedures may well be lost if the Hessian at the minimum turns out to be singular, and as we shall now show, unfortunately on the basis provided by the elements of $\boldsymbol{Y}$ the Hessian at the minimum is indeed singular.

To determine the Hessian of the problem we must find the second variation in E, and it is easy to see that under a variation $\boldsymbol{Y}\rightarrow\boldsymbol{Y}+\delta\boldsymbol{Y}$ we get $\boldsymbol{V}\rightarrow\boldsymbol{V}+\boldsymbol{V}^1$

where:

$$\begin{aligned} \boldsymbol{V}^1 &= 4\{-\boldsymbol{S}\delta\boldsymbol{R}\boldsymbol{f}\boldsymbol{Y}\boldsymbol{A}^{-1} + (\boldsymbol{I}-\boldsymbol{S}\boldsymbol{R})(\boldsymbol{f}\delta\boldsymbol{Y}\boldsymbol{A}^{-1} - \boldsymbol{f}\boldsymbol{Y}\boldsymbol{A}^{-1}\delta\boldsymbol{A}\boldsymbol{A}^{-1}) \\ &\quad + (\boldsymbol{I}-\boldsymbol{S}\boldsymbol{R})\boldsymbol{G}(\delta\boldsymbol{R})\,\boldsymbol{Y}\boldsymbol{A}^{-1}\} \end{aligned} \tag{3.15}$$

$$\begin{aligned} &= 4\Big\{(\boldsymbol{I}-\boldsymbol{S}\boldsymbol{R})(\boldsymbol{f}\delta\boldsymbol{Y}\boldsymbol{A}^{-1} - \boldsymbol{f}\boldsymbol{Y}\boldsymbol{A}^{-1}\delta\boldsymbol{A}\boldsymbol{A}^{-1}), \\ &\quad -\boldsymbol{S}(\boldsymbol{Y}\boldsymbol{A}^{-1}\delta\boldsymbol{Y}(\boldsymbol{I}-\boldsymbol{S}\boldsymbol{R}) + (\boldsymbol{I}-\boldsymbol{R}\boldsymbol{S})\delta\boldsymbol{Y}\boldsymbol{A}^{-1}\boldsymbol{Y})\boldsymbol{f}\boldsymbol{Y}\boldsymbol{A}^{-1} \\ &\quad + \sum_{j=1}^{m}\sum_{r=1}^{n}\boldsymbol{X}^{jr}\delta\boldsymbol{Y}_{jr}\Big\} \end{aligned} \tag{3.16}$$

where

$$X_{is}^{jr} = \sum_{klqp=1} (\boldsymbol{I}-\boldsymbol{S}\boldsymbol{R})_{iq}(\boldsymbol{I}-\boldsymbol{S}\boldsymbol{R})_{jp}(B_{pl,qk} + B_{lp,qk})(\boldsymbol{Y}\boldsymbol{A}^{-1})_{lr}(\boldsymbol{Y}\boldsymbol{A}^{-1})_{ks} \tag{3.17}$$

where $B_{pl,qk}$ is the electron-repulsion supermatrix with elements

$$B_{pl,qk} = 2\langle qp|g|kl\rangle - \langle qp|g|lk\rangle \tag{3.18}$$

where the integral notation is that of McWeeny and Sutcliffe [11].

Since we are interested only in the Hessian at the minimum, we can use the fact that at the minimum $\boldsymbol{V}=\boldsymbol{0}$, to to simplify (3.16) somewhat, and it is easily seen that at the minimum the second and third terms in (3.16) vanish to give

$$\begin{aligned} \boldsymbol{V}^1 = 4\Big\{&(\boldsymbol{I}-\boldsymbol{S}\boldsymbol{R})(\boldsymbol{f}\delta\boldsymbol{Y}\boldsymbol{A}^{-1} - \boldsymbol{S}\delta\boldsymbol{Y}\boldsymbol{A}^{-1}\boldsymbol{Y}\boldsymbol{f}\boldsymbol{Y}\boldsymbol{A}^{-1}) \\ &+ \sum_{j=1}^{m}\sum_{r=1}^{n}\boldsymbol{X}^{jr}\delta\boldsymbol{Y}_{jr}\Big\} \end{aligned} \tag{3.19}$$

that is

$$\begin{aligned} \frac{\partial^2 E}{\partial Y_{jr}\partial Y_{is}} = 4\{&((\boldsymbol{I}-\boldsymbol{S}\boldsymbol{R})\,\boldsymbol{f})_{ij}\boldsymbol{A}_{rs}^{-1} - ((\boldsymbol{I}-\boldsymbol{S}\boldsymbol{R})\boldsymbol{S})_{ij} \\ &\cdot(\boldsymbol{A}^{-1}\boldsymbol{Y}^T\boldsymbol{f}\boldsymbol{Y}\boldsymbol{A}^{-1})_{rs} + \boldsymbol{X}_{is}^{jr}\} \end{aligned} \tag{3.20}$$

so that the Hessian at the minimum has elements given by (3.20) where it is understood that all quantities dependent on $\boldsymbol{Y}$ in (3.20) are given in terms of the minimizing $\boldsymbol{Y}$, though this is not explicitly indicated in the equation.

From (3.14) it is easy to see that at the minimum

$$\boldsymbol{f}\boldsymbol{R}\boldsymbol{S} - \boldsymbol{S}\boldsymbol{R}\boldsymbol{f}\boldsymbol{R}\boldsymbol{S} = \boldsymbol{0} \tag{3.21}$$

and that

$$\boldsymbol{S}\boldsymbol{R}\boldsymbol{f} - \boldsymbol{S}\boldsymbol{R}\boldsymbol{f}\boldsymbol{R}\boldsymbol{S} = \boldsymbol{0} \tag{3.22}$$

so that the matrix $(\boldsymbol{I}-\boldsymbol{S}\boldsymbol{R})\boldsymbol{f}$ is symmetric at the minimum and hence the Hessian itself is symmetric, as required. We can therefore write the Hessian as a partitioned matrix of dimension n by n in blocks of dimension m by m, the MO's labelling the blocks and the AO's labelling the runs and columns within a block.

The r, s block clearly has the structure

$$\boldsymbol{H}^{rs} = 4\boldsymbol{A}_{rs}^{-1}(\boldsymbol{I}-\boldsymbol{SR})\boldsymbol{f} - 4\bar{\boldsymbol{A}}_{rs}^{-1}(\boldsymbol{I}-\boldsymbol{SR})\boldsymbol{S} + 4\boldsymbol{Z}^{rs} \tag{3.23}$$

where

$$\bar{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{Y}^T \boldsymbol{f} \boldsymbol{Y})^{-1}\boldsymbol{A} \tag{3.24}$$

and

$$Z_{ji}^{rs} = X_{is}^{jr}\,. \tag{3.25}$$

Now let us suppose that we have found a matrix $\boldsymbol{T}$, that minimises E by satisfying the usual equations

$$\boldsymbol{fT} = \boldsymbol{STe}\,, \tag{3.26}$$

$$\boldsymbol{T}^T\boldsymbol{ST} = \boldsymbol{I}\,. \tag{3.27}$$

Then we know that we can write the minimizing $\boldsymbol{R}$ as $\boldsymbol{TT}^T$ and the minimum $\boldsymbol{f}$ as the one obtained from (3.26). Consider now the mn by 1 column matrix $\boldsymbol{t}$ whose first m rows are $\boldsymbol{T}_1$ whose second m rows are $\boldsymbol{T}_2$, and so on, where $\boldsymbol{T}_r$ is the r'th column of $\boldsymbol{T}$. We can then construct

$$\hat{\boldsymbol{t}} = \boldsymbol{Ht} \tag{3.28}$$

where $\hat{\boldsymbol{t}}$ is a column matrix whose first m rows are

$$\hat{\boldsymbol{t}}^1 = \sum_{s=1}^{n} \boldsymbol{H}^{1s}\boldsymbol{T}_s \tag{3.29}$$

and so on. If we write out the expression for $\boldsymbol{t}^r$ explicitly we get

$$\boldsymbol{t}^r = 4\sum_{s=1}^{n} \boldsymbol{A}_{rs}^{-1}(\boldsymbol{I}-\boldsymbol{SR})\boldsymbol{fT} - \boldsymbol{A}_{rs}^{-1}(\boldsymbol{I}-\boldsymbol{SR})\boldsymbol{ST} + \boldsymbol{Z}^{rs}\boldsymbol{T}_s\,. \tag{3.30}$$

But

$$\begin{aligned}(\boldsymbol{I}-\boldsymbol{SR})\boldsymbol{fT}_s &= \boldsymbol{fT}_s - \boldsymbol{STT}^+\boldsymbol{fT}_s \\ &= \boldsymbol{fT}_s - \varepsilon_s\boldsymbol{ST}_s \\ &= \boldsymbol{0}\,,\end{aligned} \tag{3.31}$$

$$\begin{aligned}(\boldsymbol{I}-\boldsymbol{SR})\boldsymbol{ST}_s &= \boldsymbol{ST}_s - \boldsymbol{STTST}_s \\ &= \boldsymbol{ST}_s - \boldsymbol{ST}_s \\ &= \boldsymbol{0}\,,\end{aligned} \tag{3.32}$$

and

$$\begin{aligned}(\boldsymbol{Z}^{rs}\boldsymbol{T}_s)_j &= \sum_{i=1}^{m} Z_{ji}^{rs}T_{is} \\ &= \sum_{ip=1}^{m} X_{rs,jp}(\boldsymbol{I}-\boldsymbol{SR})_{ip}T_{is}\end{aligned} \tag{3.33}$$

where

$$X_{rs,jp} = \sum_{k,l,q=1}^{m} (\boldsymbol{I}-\boldsymbol{SR})_{jq}(B_{pl,qk}+B_{pl,kq})(\boldsymbol{YA}^{-1})_{lr}(\boldsymbol{YA}^{-1})_{ks} \tag{3.34}$$

so that

$$
\begin{aligned}
(\boldsymbol{Z}^{rs}\boldsymbol{T}_s)_j &= \sum_{i,p=1}^{m} X_{rs,jp}\left(\delta_{ip}T_{is} - \sum_{1} R_{pl}S_{li}T_{is}\right) \\
&= \sum_{p=1}^{m} X_{rs,jp}\left(T_{ps} - \sum_{il=1}^{m}\sum_{u=1}^{n} T_{pu}T_{lu}S_{li}T_{is}\right) \\
&= \sum_{p=1}^{m} X_{rs,jp}\left(T_{ps} - \sum_{u=1}^{n}\delta_{us}T_{pu}\right) \\
&= 0\,.
\end{aligned}
\tag{3.35}
$$

We therefore conclude that $\hat{\boldsymbol{t}}^r = \boldsymbol{0}$, so that we can write:

$$
\boldsymbol{H}\boldsymbol{t} = 0\boldsymbol{t} \tag{3.36}
$$

and hence we conclude that $\boldsymbol{t}$ is an eigenvector of $\boldsymbol{H}$ with zero eigenvalue so that $\boldsymbol{H}$ is singular and not positive definite.

In fact the above demonstration may easily be extended to show that the Hessian at the minimum has precisely n^2 zero roots, by the following means. We can regard the Hessian as being defined in an mn dimensional vector space and we can define a basis in this space by choosing n^2 vectors $\boldsymbol{t}_p$, $p = 1, 2, \ldots, mn$, according to the following specification. Select one column $\boldsymbol{T}_q$ from the n possible columns of $\boldsymbol{T}$. Let $\boldsymbol{t}_p$ be the vector that has $\boldsymbol{T}_q$ as its first m rows and is null elsewhere, let $\boldsymbol{t}_{p+1}$ have $\boldsymbol{T}_q$ as the second m rows and be null elsewhere and so on. It is easy to see that the n^2 vectors so chosen are linearly independent since they are orthonormal in a metric specified by the matrix partitioned as is $\boldsymbol{H}$ but with $\boldsymbol{S}$ forming the diagonal blocks and with null blocks elsewhere. It follows at once from (3.29) to (3.34) that these n^2 vectors are eigen-vectors of $\boldsymbol{H}$ with null eigen-values. The vector $\boldsymbol{t}$ that we chose in equation (3.28) is of course just the linear combination of degenerate eigenvectors

$$
\boldsymbol{t} = \boldsymbol{t}_1 + \boldsymbol{t}_{n+2} + \boldsymbol{t}_{2n+3} + \cdots + \boldsymbol{t}_{n^2}\,. \tag{3.37}
$$

4. The Origin of the Zero Roots in the Hessian

Let us suppose for the moment that we had formulated our energy expression originally in terms of a set of non-orthonormal orbitals related to the atomic basis η by the matrix $\boldsymbol{Y}$. Then it is easy to see, using the formulae for matrix elements between determinants of non-orthogonal orbitals (see e.g. [11], p. 49–51) that the energy expression obtained is just (3.1) but now with $\boldsymbol{R}$ defined directly by (3.6). Thus had we worked without any constraints at all, we would have obtained precisely the same equations as we have already for the gradient and for the Hessian and would, in consequence, have encountered precisely the same difficulties. In the light of this it is perhaps misleading to regard $\boldsymbol{U}$ in Eq. (3.4) as a constraint supplying matrix, but rather as a constraint removing matrix. One can therefore regard the minimization problem we have so far formulated as the one of determining the non-orthonormal molecular orbital at any stage

and then, using the freedom that one has on the one-determinant approximation, performing a linear transformation among them to produce an orthonormal set. It would seem likely therefore that it is precisely because we have, even at the minimum, this freedom to perform an arbitrary n by n linear transformation among the solution vectors that we have a Hessian with n^2 degenerate zero roots. It further seems likely that if instead of removing the constraints we had used the constraints to remove variables from the problem and hence effectively to remove the freedom to perform an arbitrary linear transformation among the solution vectors, then we should not have a singular Hessian for the problem. That this removal is, in certain circumstances, possible and does indeed have the required results is easy to see from the following example.

Let us suppose that we are working, not in an arbitrary basis of AO's, but in a basis of exact canonical MO's, numbered in order of increasing orbital energy $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_r, \ldots, \varepsilon_n$, and let us further suppose that we require the orbitals at any stage to be orthonormal. At the minimum we shall have a matrix

$$\boldsymbol{T} = \begin{matrix} n \\ m-n \end{matrix}\left(\begin{matrix} \boldsymbol{Q}^T \\ \hline \boldsymbol{0} \end{matrix}\right) = \boldsymbol{Y}\boldsymbol{U} \tag{4.1}$$

where $\boldsymbol{Q}$ is an arbitrary orthogonal matrix and if the minimization process happens to yield the canonical molecular orbitals then $\boldsymbol{Q}$ will be the unit matrix.

In the canonical MO basis we shall have:

$$\boldsymbol{S} = \boldsymbol{I}$$

$$f_{ij} = \varepsilon_i \delta_{ij}$$

and at the minimum

$$R_{ij} = \begin{cases} \delta_{ij};\ i, j \leqq n \\ 0;\ i, j > n \end{cases} \tag{4.2}$$

and after a little manipulation we can re-write (3.20) as

$$\begin{aligned} \frac{\partial^2 E}{\partial Y_{jr} \partial Y_{is}} = & \ 0; \quad i \text{ or } j \leqq n, \quad \text{all } r, s \\ & 4\delta_{ij}(\varepsilon_j A_{rs}^{-1}) - (\bar{\boldsymbol{U}} \boldsymbol{\varepsilon}^0 \boldsymbol{U}^T)_{rs} \\ & + \sum_{vt=1}^{n} (B_{jt;jv} + B_{tj,jv})\, \bar{U}_{vs} \bar{U}_{ts}; \quad \text{with } i \text{ and } j > n \end{aligned} \tag{4.3}$$

where

$$\bar{\boldsymbol{U}} = \boldsymbol{U}\boldsymbol{Q} \tag{4.4}$$

and where $\boldsymbol{\varepsilon}^0$ is the diagonal matrix composed of $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$.

If it so happens that the minimization process yields the canonical orbitals with $\boldsymbol{U} = \boldsymbol{I}$ also, that is of $Y_{ir} = T_{ir}$, then (4.3) simplifies further so that the Hessian has elements:

$$\begin{aligned} \frac{\partial^2 E}{\partial Y_{jr} \partial Y_{is}} = & \ 0; \quad i, j \leqq n \\ = & \ [4(\varepsilon_j - \varepsilon_r)\, \delta_{ij} \delta_{rs} \\ & + 8\langle ij|g|sr\rangle - 4\langle ij|g|rs\rangle; \quad i, j > n \\ & + 8\langle ir|g|sj\rangle - 4\langle ir|g|js\rangle]. \end{aligned} \tag{4.5}$$

This matrix may easily be transformed by elementary operations to be a matrix null except for one diagonal block of side $mn-n^2$, hence confirming that the Hessian has just n^2 zero roots.

Now let us partition $\boldsymbol{Y}$ as $\boldsymbol{T}$ is partitioned in (4.1) viz:

$$\boldsymbol{Y}=\begin{matrix} n \\ m-n \end{matrix}\left(\begin{matrix} \boldsymbol{Y}_t \\ \hline \boldsymbol{Y}_b \end{matrix}\right) \tag{4.6}$$

thus from (4.1), at the minimum we must have

$$\boldsymbol{Y}_t\boldsymbol{U}=\boldsymbol{Q}^T, \tag{4.7}$$

$$\boldsymbol{Y}_b\boldsymbol{U}=\boldsymbol{0}. \tag{4.8}$$

Assuming that $\boldsymbol{U}$ is non-singular (4.8) implies that at the minimum $\boldsymbol{Y}_b$ is a null matrix, and therefore at the minimum the orthogonality constraint (3.3) becomes

$$\boldsymbol{U}^T\boldsymbol{Y}_t^T\boldsymbol{Y}_t\boldsymbol{U}=(\boldsymbol{Y}_t\boldsymbol{U})^T(\boldsymbol{Y}_t\boldsymbol{U})=\boldsymbol{I} \tag{4.9}$$

thus at the minimum the matrix $\boldsymbol{Y}_t\boldsymbol{U}$ must, by virtue of the orthogonality constraint alone, be an orthogonal matrix. This means that we can make an arbitrary initial choice of $\boldsymbol{Y}_t$, and keep this choice fixed throughout the minimization as long as $\boldsymbol{U}$ remains defined and a minimum can be found with this choice of $\boldsymbol{Y}_t$. It is obvious in this case that a minimum can be found if we choose initially $\boldsymbol{Y}_t$ as the unit matrix (or indeed as any orthogonal matrix) and make an initial estimate of $\boldsymbol{Y}_b$ such that $\boldsymbol{Y}_b^T\boldsymbol{Y}_b$ has all its eigenvalues greater than -1. In fixing $\boldsymbol{Y}_t$ we have removed n^2 variables from the problem and thus the gradient matrix now has $mn-n^2$ elements and the Hessian matrix consists of blocks $\boldsymbol{H}^{rs}$ composed simply of the non-zero parts of (4.3). It is easy to see that with this choice all the singularities in the Hessian have been eliminated. Indeed it is fairly clear that one need not (though one may on this basis) fix all n^2 elements of $\boldsymbol{Y}_t$. It is sufficient to fix $\frac{1}{2}n(n+1)$ of them, say by choosing $\boldsymbol{Y}_t$ initially to be a unit upper triangle, and the singularities in the Hessian will still be removed and a solution will still be possible.

In the case where we work in an arbitrary basis the situation is not so clear however precisely because we cannot say which elements of $\boldsymbol{Y}$ we may fix and still obtain a minimizing solution. Indeed it is not clear even in the very simple case in which $\boldsymbol{Y}$ consists of a single column only. In this case the problem is very similar to the problem of finding the lowest eigenvalue of a real symmetric matrix by minimizing the Rayleigh quotient, a problem which is discussed by Fletcher and Bradbury [13]. Here one can obviously avoid the singularity by fixing one element of the column at 1. This effectively confines the possible solutions of the problem to these lying on the faces of hypercube and the face of the hypercube along which minimization is to occur is determined by the choice of element set equal to 1. However it is not known in advance whether a minimum exists on this face. If a minimum does not exist on the chosen face then this should be shown up by an element other than that chosen becoming greater than 1, in which case the vector must be renormalized so that the emerging element is now 1, this

element must then be fixed and the minimization process started afresh along the newly chosen face.

In the general case the equivalent tactic would be to fix $\frac{1}{2}n(n+1)$ values of $\boldsymbol{Y}$ and then to attempt to minimize the energy. Presumably the equivalent behaviour in the event of no minimum existing would be for the matrix $\boldsymbol{A}$ to become singular. If this occured then one would simply have to choose a new set of fixed values that avoided this. Alternatively, of course, there is the possibility that fixing elements of $\boldsymbol{Y}$ makes the closest stationary point of the function, one other than the lowest minimum sought (an "excited state"). In this case also one would have simply to restart the process from with a new choice of fixed elements.

It is clearly not possible to specify any general strategy in respect of choosing the elements of $\boldsymbol{Y}$ to be fixed, beyond saying that of course no more than n elements and no less than one element should be fixed in any one column. It seems a case where numerical trials alone can decide whether a general strategy is possible for the problem. In this context it is worthwhile remembering that numerical experience up to now does indicate that the Fletcher-Reeves method, for example, converges quite well, even allowing all the elements of $\boldsymbol{Y}$ to vary, until one is fairly close to the minimum. A sensible strategy might therefore be to treat the full problem for a number of iterations and then to inspect the elements of the resulting $\boldsymbol{T}$ matrix and fix $\frac{1}{2}n(n+1)$ of these in some plausible manner to define a starting point for the minimization in the space of reduced dimersion.

It should also be noticed here that it is of course possible to reformulate our problem completely in terms of $mn-\frac{1}{2}n(n+1)$ independent variables and a way of doing this has been developed, for example by Raffinetti and Ruedenberg [14]. If such a course is adopted then it seems clear that the Hessian in the basis of independent variables should be free of singularities and thus a direct method cast in terms of these variables, should work efficiently. However it would seem to be the case that it is very difficult to get analytic expressions for the gradient matrix (and even more difficult to obtain the Hessian) in such bases. This difficulty effectively confines the choice of a direct method to a non-gradient method (such as Powell's conjugate direction method). Such methods are generally found to be slower than gradient methods, so that it is not clear that the formulation in terms of independent variables should lead to a superior convergence rate in practice and whether it is more effective to use a $\boldsymbol{Y}$ matrix formulation or not therefore remains an open question. Our analysis so far however does suggest another method of tackling this problem in which we can actually utilize the presence of zero roots in the Hessian.

5. A Direct Method Which Avoids the Zero Root Problem

Suppose that at any stage of a minimization process we have a matrix

$$\boldsymbol{T}_1 = \boldsymbol{Y}\boldsymbol{U}, \qquad \boldsymbol{T}_1^T \boldsymbol{S} \boldsymbol{T}_1 = \boldsymbol{I}_n \tag{5.1}$$

and from the matrix $\boldsymbol{f}$ evaluated at $\boldsymbol{T}_1$ we construct a matrix $\bar{\boldsymbol{f}}$ such that

$$\bar{\boldsymbol{f}} = \boldsymbol{T}_1^T \boldsymbol{f} \boldsymbol{T}_1 \tag{5.2}$$

and find a matrix $\boldsymbol{Q}_1$ which diagonalizes $\bar{f}$ such that

$$\bar{f}\boldsymbol{Q}_1 = \boldsymbol{Q}\varepsilon_1, \quad \boldsymbol{Q}_1^T\boldsymbol{Q}_1 = \boldsymbol{I}_n \tag{5.3}$$

and define a matrix $\boldsymbol{Z}_1 = \boldsymbol{T}_1\boldsymbol{Q}_1$. Let us also invent an m by $m-n$ matrix $\boldsymbol{Z}_2$ and define a new basis

$$\phi = (\phi_1 \mid \phi_2) = \eta(\boldsymbol{Z}_1 \mid \boldsymbol{Z}_2) \tag{5.4}$$

such that in this basis

$$f \to \bar{f} = \left(\begin{array}{c|c} \varepsilon_1 & \bar{f}_{12} \\ \hline \bar{f}_{21} & \varepsilon_2 \end{array}\right), \quad \bar{f}_{ij} = \boldsymbol{Z}_i^T f \boldsymbol{Z}_j \tag{5.5}$$

where ε_2 is diagonal

$$\boldsymbol{S} \to \bar{\boldsymbol{S}} = \boldsymbol{I}_m \tag{5.6}$$

and in this basis clearly

$$\boldsymbol{T}_1 \to \bar{\boldsymbol{T}} = \left(\begin{array}{c} \boldsymbol{I}_n \\ \hline \boldsymbol{0} \end{array}\right). \tag{5.7}$$

It then follows immediately from (3.14) (5.5), (5.6), and (5.7) that the gradient matrix with respect to the elements of T is just

$$\boldsymbol{V} = 4\left(\begin{array}{c} \boldsymbol{Q} \\ \hline \bar{f}_{21} \end{array}\right). \tag{5.8}$$

We can find the Hessian at the point $\bar{\boldsymbol{T}}$ (that is, *not* at the minimum) by noting from (3.15) that in general we must add to (3.20) a term for the (jr, is) element

$$-(\boldsymbol{S}\boldsymbol{Y}\boldsymbol{A}^{-1})_{ir}V_{js} + (\boldsymbol{S}\boldsymbol{Y}\boldsymbol{A}^{-1})_{js}V_{ir} + (\boldsymbol{V}\boldsymbol{Y}^T\boldsymbol{S})_{ij}(\boldsymbol{A}^{-1})_{rs} \tag{5.9}$$

where $\boldsymbol{V}$ is given by (3.14).

After some manipulation it can be shown that the Hessian with respect to the elements of $\bar{\boldsymbol{T}}$ at $\bar{\boldsymbol{T}}$ is blocked [cf. Eq. (3.23)] with blocks of the form

$$\boldsymbol{H}^{rs} = 4\delta_{rs}\left(\begin{array}{c|c} \boldsymbol{0} & \boldsymbol{B}_{rr}^{(1)} \\ \hline \boldsymbol{B}_{rr}^{(2)} & \boldsymbol{H}_r \end{array}\right) + 4\left(\begin{array}{c|c} \boldsymbol{0} & \boldsymbol{B}_{rs}^{(1)} \\ \hline \boldsymbol{B}_{rs}^{(2)} & \boldsymbol{H}_{rs} \end{array}\right). \tag{5.10}$$

Where $\boldsymbol{H}_r$ consists of the non-zero diagonal terms from (4.5):

$$(H_r)_i = (\varepsilon_{n+i} - \varepsilon_r) + 3\langle (n+i)\, r \mid g \mid r(n+i)\rangle - \langle (n+i)\, r \mid g \mid (n+i)\, r\rangle \tag{5.11}$$

and the matrix $\boldsymbol{H}_{rs}$ consists of the equivalent off-diagonal terms.

The matrix $\boldsymbol{B}_{rs}^{(1)}$ is null but for its r'th row which is the s'th row of $\bar{f}_{12}$ and $\boldsymbol{B}_{rs}^{(2)}$ is similarly null except for its s'th column which is the r'th column of $\bar{f}_{21}$.

Now we have seen in Eq. (2.10) that it is always possible to decrease the value of locally quadratic function if we choose a direction $\boldsymbol{p}$ such that $\boldsymbol{g}^T\boldsymbol{p} \neq 0$ and $\boldsymbol{p}^T\boldsymbol{H}\boldsymbol{p} > 0$, even if $\boldsymbol{H}$ is not positive definite. Now if we choose our direction vector in this problem so that its r'th group of m rows is given by

$$\boldsymbol{p}_r = -\frac{1}{4}\left(\begin{array}{c|c} \boldsymbol{0} & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{H}_r^{-1} \end{array}\right)\boldsymbol{V}_r \tag{5.12}$$

where $\boldsymbol{V}_r$ is the r'th column of $\boldsymbol{V}$, then it follows at once that

$$\boldsymbol{g}^T \boldsymbol{p} = \boldsymbol{p}^T \boldsymbol{H} \boldsymbol{p} = -4 \sum_{r=1}^{m} \sum_{j=n+1}^{m} \bar{f}_{jr} (\boldsymbol{H}_r^{-1}) \bar{f}_{jr} \tag{5.13}$$

and since the elements of $\boldsymbol{H}_r$ are practically certain to be positive we have achieved the desired result. It is interesting to note as an aside that (5.13) also implies that the optimum step length, α, along $\boldsymbol{p}$ is unity and that at a true minimum even though $\bar{f}_{12} = 0$, no divergence is encountered in the descent formulae (2.10) and (2.11).

We can re-write (5.12) in terms of a rectangular m by r matrix $\boldsymbol{P}$ with elements

$$P_{ir} = \begin{matrix} 0, & i \leqq n \\ -(\bar{f}_{21})_{i-n,r} / [(\varepsilon_i - \varepsilon_r) + 3\langle ir|g|ri\rangle - \langle ir|g|ir\rangle] \end{matrix} \tag{5.14}$$

and then the next point in the descent is found by constructing

$$\hat{\bar{\boldsymbol{T}}} = \bar{\boldsymbol{T}} + \lambda \boldsymbol{P}$$

and minimizing the energy with respect to λ. Using (5.4) we can write the change in terms of the $\boldsymbol{Z}_i$ as

$$\hat{\boldsymbol{Z}}_1 = \boldsymbol{Z}_1 + \lambda \boldsymbol{Z}_2 \boldsymbol{P}_b \tag{5.15}$$

where $\boldsymbol{P}_b$ is just the non-zero part of $\boldsymbol{P}$ written as an $(m-n)$ by n matrix.

The matrix $\boldsymbol{Z}_1$ does not of course satisfy the orthonormality constraints, so it should be regarded as a next estimate of $\boldsymbol{Y}$ and treated accordingly to determine λ. It violates the constraints only by terms of order $(P_{ir})^2$ which vanish as the minimum is reached.

The up-dating Eq. (5.15) is seen at once to be of precisely the same form as that proposed by Hillier and Saunders [15] and indeed our up-dating matrix $\boldsymbol{P}_b$ differs from the up-dating matrix $\boldsymbol{B}$ proposed by these authors only in the presence in $\boldsymbol{P}_b$ of electron interaction terms in the denominator. Since these terms are probably in most instances small compared with the orbital energy differences, it is perhaps legitimate to neglect them and we can regard this method as being essentially equivalent to that of Hillier and Saunders, but obtained from different considerations. The implementation scheme that we would propose for the method is exactly that proposed by Hillier and Saunders except that we would recommend a numerical search to minimize the energy with respect to λ rather than the analytical method described by these authors since by this means orthogonality among the orbitals can be vigorously maintained at every stage.

Finally it is interesting to note that one can make a much more simple assumption about the structure of $\boldsymbol{P}$ by replacing $\boldsymbol{H}_r^{-1}$ in (5.12) by a multiple of the unit matrix and this still preserves the local descent properties of the method. The method in this case becomes a variant of steepest descents and in this case (5.18) is replaced by

$$\hat{\bar{\boldsymbol{T}}} = \bar{\boldsymbol{T}} - \lambda \bar{\boldsymbol{V}} \tag{5.16}$$

and (5.19) becomes

$$\begin{aligned}\hat{\boldsymbol{Z}}_1 &= \boldsymbol{Z}_1 - 4\lambda \boldsymbol{Z}_2 \bar{\boldsymbol{f}}_{21} \\ &= \boldsymbol{Z}_1 - \lambda \boldsymbol{Z}_2 \boldsymbol{Z}_2^T \boldsymbol{f} \boldsymbol{Z}_1\end{aligned} \tag{5.17}$$

and using the resolution of the identity we see that

$$\boldsymbol{Z}_2 \boldsymbol{Z}_2^T = \boldsymbol{S}^{-1} - \boldsymbol{R} \tag{5.18}$$

and therefore

$$\boldsymbol{Z}_1 = \boldsymbol{Z}_1 - 4\lambda [\boldsymbol{S}^{-1} - \boldsymbol{R}] \boldsymbol{f} \boldsymbol{Z}_1 . \tag{5.19}$$

Obviously (5.19) is a much simpler formula to implement than (5.15) and as such might well be worth some numerical investigation even though steepest descent methods are not generally considered to be very effective.

6. Summary

We have demonstrated in this article that the Hessian matrix at the minimum with respect to the linear coefficients can be singular in the closed shell LCAO-MO-SCF problem, for a particular and highly convenient way of choosing the coefficients. We suggest that this is the reason why only slow convergence has been achieved using such methods as the Fletcher-Reeves method, in minimizing the LCAO-MO energy. The generalization of our results to the UHF problem is immediate and obvious.

We have suggested a method by which the singularities may be avoided by constraining the coefficients and which may well therefore have superior convergence properties. We have also established the status of the method of Hillier and Saunders among direct methods as a quasi-Newton method and suggested a new steepest descent method in the spirit of Hillier and Saunders' method.

It would seem likely from our analysis that the singularities in the Hessian arise because of the freedom one has in the closed shell case to perform a linear transformation among the occupied orbitals while leaving the energy invariant. One may perhaps conjecture that such singularities may well arise in all problems where one has such a freedom but not in problems where this freedom is absent.

References

1. McWeeny, R.: Proc. Roy. Soc. A **235**, 496 (1956)
2. McWeeny, R.: Rev. Phys. **32**, 335 (1960)
3. Fletcher, R.: Mol. Phys. **19**, 55 (1970)
4. Kari, R., Sutcliffe, B.T.: Chem. Phys. Letters **7**, 149 (1970)
5. Claxton, T., Smith, W.: Theoret. Chim. Acta (Berl.) **22**, 399 (1971)
6. Fletcher, R., Reeves, C.M.: Comput. J. **7**, 149 (1964)
7. Powell, M.J.D.: Comput. J. **7**, 155 (1964)
8. Fletcher, R., Powell, M.J.D.: Comput. J. **6**, 163 (1963)
9. Huang, H.Y.: J. Opt. Th. Appl. **5**, 405 (1970)
10. Dixon, L.C.W.: Math. Programming **2**, 383 (1972)

11. McWeeny, R., Sutcliffe, B. T.: Methods of molecular quantum mechanics. London, New York: Academic Press 1969
12. Kari, R., Sutcliffe, B. T.: Int. J. Quant. Chem. **7**, 459 (1973)
13. Bradbury, W., Fletcher, R.: Num. Math. **9**, 259 (1966)
14. Raffenetti, R., Ruedenberg, K.: Int. J. Quant. Chem. **35**, 625 (1971)
15. Hillier, I., Saunders, V.: Proc. Roy. Soc. A **320**, 161 (1970/71)

Dr. B. T. Sutcliffe
Department of Chemistry
University of York
York Y01 5DD/England