A BROYDEN–FLETCHER–GOLDFARB–SHANNO OPTIMIZATION PROCEDURE FOR MOLECULAR GEOMETRIES

John D. HEAD [1] and Michael C ZERNER

*Quantum Theory Project Department of Chemistry, University of Florida Gainesville FL 32611 USA
and Guelph Waterloo Centre for Graduate Work in Chemistry, Department of Chemistry University of Guelph,
Guelph, Ontario Canada N1G 2W1*

Most quantum-chemical calculations for geometries evaluate first derivatives of the energy with respect to nuclear positions analytically and then use update procedures to build up information on the second derivatives as they step along the potential energy surface toward a minimum (stable geometry) or simple saddle point (transition state) We describe here the use of the Broyden–Fletcher–Goldfarb–Shanno (BFGS) quasi-Newton update used in conjunction with a partial line search. We have found BFGS superior to the other update formulae we have examined In particular, it is superior to the Murtagh–Sargent (MS) scheme that is commonly used in geometry determinations The advantage of the BFGS update over the MS scheme becomes especially dramatic for large molecular systems

1 Introduction

The use of molecular quantum mechanics to obtain molecular geometry has become widespread The structures of molecules as yet unmade and of excited-state molecules present few additional problems Transition states that can only be inferred from experiments are somewhat more difficult to obtain but utilize the same general techniques.

The most common techniques used to search a molecular energy hypersurface obtained either through semi-empirical or ab initio quantum mechanics require analytic first derivatives of the energy with respect to coordinate displacement, and then proceed to build up curvature information on the surface as they move along it. The most common of these methods used are of the Murtagh–Sargent (MS) [1] type first introduced into quantum chemistry by McIver and Komornicki [2] For small molecules this method has proven quite successful For larger molecules gradients are generally reduced to below 10^{-2} au/bohr with reasonable efficiency, but a further monotonic reduction of the gradients to an acceptable value is quite slow It is this difficulty that has prompted us to re-examine the various update procedures that are available and to test their performance, especially with respect to the popular MS method In doing this we have found one method, a variant of the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [3–7], that is systematically superior in obtaining stable molecular geometries, especially for large molecules. This method does reduce the gradients in a more or less monotonic fashion. We describe this technique here and compare it with our version of the MS method. A more complete report on these methods as well as others we have examined is in progress [8].

BFGS methods have been used by Simons, Jørgensen, Taylor and Ozment [9] and a version of this technique has been used by Spangler, Williams and Maggiora [10] This technique described here differs essentially in the line search algorithm.

2. Quasi-Newton methods

The energy $E(x^k)$ and the gradient vector $g(x^k)$ are expanded in a Taylor series yielding

[1] Present address Department of Chemistry University of Hawaii at Manoa, Honolulu, Hawaii, USA

$$E(x^{k+1}) = E(x^k) + g^{k+} \cdot s^k + \tfrac{1}{2} s^{k+} H^k s^k + \quad, \tag{1a}$$

$$g^{k+1} \equiv g(x^{k+1}) + H^k s^k + \ldots, \tag{1b}$$

with

$$s^k = x^{k+1} - x^k. \tag{1c}$$

Extreme points, such as those that characterize minima, saddle points or maxima, are characterized by $g = 0$. Assuming that $g^{k+1} \approx 0$ for large enough k and truncating eq (1b) after the second term, yields the familiar Newton equation for the search direction, s^k

$$s^k = -H^{-1} g^k. \tag{2}$$

For an exact H, s^k is said to be a true "Newton" step toward the extremum. A sequence of such steps determines a Newton procedure.

Since the calculation of H either analytically or numerically is a process many times slower than that for obtaining the energy and gradients, methods are used to approximate H. The most common of these are the quasi-Newton methods Since most optimization algorithms repeatedly solve eq. (2) for s^k, the inverse Hessian is usually sought *directly*. An analogous development for the Hessian itself is easily developed.

The quasi-Newton methods relate the differences of gradients obtained at two different points on the potential energy hypersurface

$$\gamma^k = g^{k+1} - g^k \tag{3}$$

to the differences in the coordinates

$$\delta^k = \alpha^k s^k = x^{k+1} - x^k, \tag{4}$$

which, for a quadratic function, should obey the relation

$$\gamma^k = H \delta^k \tag{5}$$

Quasi-Newton methods are constrained to obey the "quasi-Newton" condition obtained from rearranging eq (5),

$$G^{k+1} \gamma^k = \delta^k, \tag{6a}$$

$$G^{k+1} \approx (H^{k+1})^{-1}. \tag{6b}$$

The Murtagh—Sargent method is a rank-1 method defined by

$$G^{k+1} = G^k + \varepsilon^k = G^k + a|u\rangle\langle u|, \tag{7}$$

where we have used the common Dirac notation. Use of eq. (6a) yields

$$G^k|\gamma^k\rangle + a|u\rangle\langle u|\gamma^k\rangle = |\delta^k\rangle. \tag{8}$$

Choosing the constant a in eq. (8) as $\langle u|\gamma^k\rangle^{-1}$ and substituting into eq. (7) yields the update equations (3)–(6)

$$G^{k+1} = G^k + \frac{(|\delta^k\rangle - G^k|\gamma^k\rangle)(\langle\delta^k| - \langle\gamma^k|G^k)}{\langle\gamma^k|\delta^k - G^k\gamma^k\rangle} \tag{9}$$

that forms the bases of all rank-1 methods

The BFGS method is a rank-2 method

$$G^{k+1} = G^k + a|u\rangle\langle u| + b|v\rangle\langle v| \tag{10}$$

Various choices of $|u\rangle$ and $|v\rangle$ and of a and b distinguish these methods The BFGS choice [7] leads to

$$G^{k+1}_{\text{BFGS}} = G^k + \frac{(\langle\delta^k|\gamma^k\rangle + \langle\gamma^k|G^k|\gamma^k\rangle)|\delta^k\rangle\langle\delta^k|}{|\langle\delta^k|\gamma^k\rangle|^2}$$

$$- \frac{|\delta^k\rangle\langle\gamma^k|G^k + G^k|\gamma^k\rangle\langle\delta^k|}{\langle\delta^k|\gamma^k\rangle}. \tag{11}$$

If Cartesian coordinates are used, G^1 is usually set to the unit matrix [‡]. The first step, s^1, is then in the direction of steepest descent Subsequent cycles update G toward greater accuracy as they move along the surface.

The rank-1 formula has two principal disadvantages The positive definiteness of G^k may be lost on updating Second, the denominator of eq. (9) may approach zero sufficiently rapidly to lead to numerical inaccuracy Murtagh and Sargent have devised simple methods to detect these problems. There seems to be no completely successful solution to either of these problems when they do occur, however, and the inverse Hessian matrix is reset to the unit matrix

In contrast the BFGS formula ensures that the Hessian remains positive definite on all updates A positive definite Hessian is desirable because it always generates a search direction s^k, given by eq. (12) below, which is in the direction of energy decrease. The BFGS formula can even be advantageous over the exact Hessian H in regions where the exact Hessian has negative eigen-

---

[‡] If internal coordinates y are used, this choice of $G^1 = 1$ for Cartesian coordinates x corresponds to $G_{y,1} = (B^{-1})^+ B^{-1}$ where $y = Bx$ and $B^{-1} = (B^+ mB)^{-1} B^+ m$ where m is an arbitrary $(3N - 6) \times (3N - 6)$ matrix, see, for example, ref [11]

values A search direction will have uphill steps along any normal mode of a molecule which has a negative eigenvalue [8].

3. Line searches

With G^k updated from either the MS or BFGS formula, a search direction, s^k, is determined by

$$s^k = -G^k g^k . \tag{12}$$

A line search is then performed to determine how far along this direction a step, δ^k, should proceed,

$$x^{k+1} = x^k + \alpha^k s^k , \tag{13a}$$

$$\delta^k \equiv \alpha^k s^k . \tag{13b}$$

A weak line search requires only that $E^{k+1} < E^k$ if searching for a minimum An exact line search finds that value of α^k for which $E(x^k + \alpha^k s^k)$ is a minimum There is thus a competition between the number of energy evaluations that are required to minimize $E(x^k + \alpha^k s^k)$ along a given s^k, or accepting nearly any move along s^k that reduces the energy or the gradient norm and then finding a new direction s^{k+1} at the new geometry

We have adopted the following line search strategy [8]:

In the kth line search cycle we examine the interval (α_1^k, α_2^k) of acceptable α^k values Initially $\alpha_1^k = 0$ and α_2^k is determined by

$$|\alpha_2^k s^k| \leqslant \Delta = 0.4 \text{ au} , \tag{14}$$

a condition preventing arbitrary large coordinate moves. The value 0.4 au is itself, of course, arbitrary, and larger values might be acceptable for larger molecules The initial α^k value is set to 1 except for the first cycle where α^1 is set to 0.4 In this first cycle $G^1 = 1$ and the move is along the direction of steepest descent The value $\alpha^1 = 0.4$ is empirical and nearly always satisfies the line search conditions even for poor starting geometries

If the initial $\alpha^k > \alpha_2^k$ then α^k is reset to α_2^k $E^{k+1} = E(x^k + \alpha^k s^k)$ is evaluated and

$$E^k - E^{k+1} \geqslant -\rho \alpha^k \langle s^k | g^k \rangle \tag{15}$$

checked $\langle s^k | g^k \rangle$ is the first term in the Taylor series expansion for the energy and is negative for a decrease

in energy. From eq (1) and eq. (2) $0 \leqslant \rho \leqslant 1/2$, with $\rho = 1/2$ corresponding to an exact line search and any positive value simply leading to a decrease in energy. After considerable numerical testing, we chose $\rho = 0 01$ corresponding to a "weak" search. If this test fails, however, $\alpha^k > \alpha_2^k$(true) $\alpha_2^k \rightarrow \alpha^k$ and α^k is reduced using the information already available according to

$$\alpha^k \rightarrow \alpha^k + \tfrac{1}{2}(\alpha^k - \alpha_1^k) \left(1 + \frac{E^k - E^{k+1}}{(\alpha^k - \alpha_1^k)E'(\alpha_1^k)} \right)^{-1} , \tag{16a}$$

$$E'(\alpha_1^k) = g(x^k + \alpha_1^k s^k)^+ \cdot s^k = (\partial E/\partial \alpha)_{\alpha = \alpha_1^k} \tag{16b}$$

The energy $E^{k+1} \rightarrow E(x^k + \alpha^k s^k)$ is now computed and the right test of eq (15) repeated until the test is obeyed

The gradient $g^{k+1} = g(x^k + \alpha^k s^k)$ is now evaluated and a left extreme test applied

$$|\langle g^{k+1} | s^k \rangle| \leqslant -\sigma \langle g^k | s^k \rangle, \quad 0 \leqslant \sigma \leqslant 1 . \tag{17}$$

For an exact line search $\langle g^{k+1} | s^k \rangle = 0$ (the gradient is perpendicular to the search direction) corresponding to $\sigma = 0$. The value $\sigma = 1$ only requires a reduction of the scalar product $\langle g^{k+1} | s^k \rangle$. For the test of eq (17) we choose $\sigma = 0 9$, again corresponding to a weak line search If the left test fails, however, α^k is too small, and a new value is estimated from

$$\alpha^k(\text{new}) = \alpha^k + \frac{(\alpha^k - \alpha_1^k)E'(\alpha^k)}{E'(\alpha_1^k) - E'(\alpha^k)}. \tag{18}$$

The energy $E^k \rightarrow E(x^k + \alpha^k(\text{new}) s^k)$ is evaluated and both right and left line test repeated with $\alpha_1^k \rightarrow \alpha^k$ and $\alpha^k \rightarrow \alpha^k(\text{new})$. When both tests have been successfully passed, the partial line search is finished and a new search direction is sought

In the MS procedure the test for a positive definite Hessian may fail. If so G^k is reset to the unit matrix and the line search is conducted as if this were once again the first cycle

4. Optimization algorithms

An initial geometry x_1 is chosen. E^1 and g^1 are calculated. Then,

(i) a search direction s^k is chosen through eq. (12). Then,

(ii) a line search is performed to determine how far along this direction to move as described in previous section. Then,

(iii) a test for convergence is made If convergence is satisfactory, the procedure is complete. If not,

(iv) estimate a new Hessian or its inverse and repeat the procedure

The success of a procedure is measured in the number of such cycles that are required to find the extreme point sought. The fewer energy evaluations, the better the method

5. Results and discussion

Table 1 summarizes results of geometry optimizations comparing the MS and BFGS methods The notation $x(y, z)$ that we use indicates that x line searches, y energy evaluations and z gradient evaluations were required starting with geometries that range from quite good to very poor An indication of the quality of the starting geometry is given by the maximum value of the gradient. Values greater than 0.1 au/bohr usually indicate one or more bond lengths are displaced by more than 0 1 Å. Two of the larger molecules of this study were not successfully optimized by the MS pro-

cedure After fifty cycles, the largest gradient of $[(CH_3)_2P(O)O]^-$ still larger than 2×10^{-3} au/bohr and that for CH_2CHCHO greater than 7×10^{-3} au/bohr The former case is difficult as the two methyl groups are rotating relative to one another to an eclipsed conformation along two very soft modes

Table 2 is a typical example of how an optimization proceeds for the rank-1 update using the line search described. Line 4 of the Hessian eigenvalues indicates that the signature of the Hessian matrix has been destroyed by the third update. The negative eigenvalue causes the step s^4 to go uphill on the energy surface even though $(g|s)$ is negative There are, as a consequence, repeated failures of the right extreme line search test eq. (15) and after the eighth energy evaluation the line search terminates without a proper α value At this stage another Hessian update is taken using the gradients evaluated at the coordinates x^4 and x^8 and fortunately the new Hessian generated is positive definite. In the remaining cycles the Hessian remains positive definite and the NH_3 geometry optimizes smoothly. We note that even after all gradients have been reduced below 10^{-4} au/bohr, the approximate Hessian eigenvalues are still not accurate

The MS procedure would detect the negative eigenvalue at cycle 4 and reset the Hessian matrix to the

Table 1
Some examples comparing the MS and BFGS update schemes [a]

	max $\|g_i\|$ [b]	MS	BFGS
NH_3 [c]	0 112	7(9,8)	7(8, 8)
NH_3 [d]	0 198	7(11,8)	7(8, 8)
H_2O_2	0 024	5(5,5)	6(6, 6)
C_2H_6	0 193	17(19,18)	19(20, 20)
H_2CO	0 264	9(11,10)	11(12,12)
$H_2CCHCHO$	0 105	NO [e]	24(27,26)
$[(CH_3)_2P(O)O]^-$	0 119	NO [e]	43(44,44)
pyrimidine,			
$C_4N_2H_4$	0.127	11(14,13)	10(11,11)
[2,2']-paracyclophane			
$C_{16}H_{16}$	0 050	16(20,18) [f]	10(11,11)

[a] The notation $x(y, z)$ refers to x line searches, y energy evaluations and z gradient evaluations

[b] The largest initial gradient in au/bohr giving a measure of the "poorness" of the initial geometry. Convergence is assumed when $\max \|g_i^k\| < 10^{-4}$ au/bohr, except for $[(CH_3)_2P(O)O]^-$, pyrimidine and [2,2']-paracyclophane where $\max \|g_i^k\| < 10^{-3}$ au/bohr.

[c] Displaced 0 05 au along each normal mode   [d] Displaced 0 08 au along each normal mode

[e] Not optimized after 50 energy evaluations

[f] For paracyclophane $\max \|g^k\| < 10^{-3}$ au/bohr at 14(16, 15) but the subsequent cycle has an increase in gradient above this threshold even though the energy monotonically decreases

Table 2
Rank-1 update optimization of NH_3 — Hessian eigenvalues (au)

1	1.000D + 00	1 000D + 00	1.000D + 00	1 000D + 00	1 000D + 00	1 000D + 00
2	2.724D + 00	1 000D + 00	1 000D + 00	1.000D + 00	1 000D + 00	1 000D + 00
3	3 227D + 00	2 714D + 00	1 000D + 00	1.000D + 00	1 000D + 00	1 000D + 00
4	−2.354D − 02	2 717D + 00	1 802D + 00	1 000D + 00	1 000D + 00	1 000D + 00
8	2 719D + 00	2.331D + 00	1 458D + 00	1.000D + 00	1 000D + 00	1 689D − 01
9	2 587D + 00	2 303D + 00	1 456D + 00	1 000D + 00	9 994D − 01	1 581D − 01
10	2 370D + 00	2 267D + 00	1 438D + 00	9 999D − 01	9 862D − 01	1 552D − 01
11	2 359D + 00	2 262D + 00	1 435D + 00	9.997D − 01	5 302D − 01	1.506D − 01
exact	2 2744D + 00	2 2744D + 00	1 4519D + 00	1 5998D − 01	1 5430D − 01	1 5430D − 01

Evaluation number		Energy	$|g|$	$\max|g_i|$	α
E	G				
1	1	−12 499372	0 358164	0.198326	0 00000D + 00
2	2	−12.521284	0 050917	0 031968	4 00000D − 01
3	3	−12 522080	0 028010	0 017822	1.00000D + 00
4	4	−12 522281	0 019552	0 012914	1 00000D + 00
5		−12 476193			8 67976D − 01
6		−12 520752			8 01833D − 02
7		−12.522073			1 55110D − 02
8	5	−12 522235	0 020010	0 012987	3 67105D − 03
9	6	−12 523207	0 004431	0.002909	1.00000D + 00
10	7	−12 523216	0.000612	0 000343	1 00000D + 00
11	8	−12 523216	0.000086	0 000049	1.00000D + 00

unit matrix and $\alpha^4 = 0.4$ In spite of the loss of all previous Hessian information by the MS procedure, one *less* cycle is required to complete the optimization In $[(CH_3)_2P(O)O]^-$ the Hessian is reset 17 times in 50 cycles, and there seems to be no more successful procedure for controlling the rank-1 update than that suggested by Murtagh and Sargent. A somewhat more successful procedure of the MS type, however, does result by setting the right hand extreme, $\alpha_2 = 1$; that is, never permitting an α^k value greater than 1

Table 3 examines this same situation for the BFGS method. As expected all the eigenvalues of the approximate Hessian remain positive on update, and the optimization proceeds rather smoothly At the second cycle two eigenvalues of the Hessian are upgraded, a characteristic of the rank-2 nature of this method, to be compared to the one value upgraded in the rank-1 schemes In spite of this, at convergence the Hessian eigenvalues obtained from the BFGS scheme are no better than those obtained from the MS procedure

For larger molecules both the MS and the BFGS method seem equally able to reduce the gradients to below about 10^{-2} au/bohr, usually corresponding to

bond lengths to within ± 0 02 Å and bond angles ± 2–$3°$ of their final values Dihedral angles, however, usually having to do with soft torsional or umbrella-like modes, are not accurate with this level of convergence It is here that the BFGS method is rather superior. This is indicated in table 4 for $H_2CCHCHO$ MS1 is the procedure described in this work, MS2 is a Murtagh–Sargent procedure in which $\alpha_2 = 1$ as briefly discussed above This table is typical in that it shows the erratic behavior of the gradient norm $|g|$ with the cycling for the MS procedure. To cycle 10 the methods have reasonably similar behavior. After that, however, the BFGS method is clearly superior. By cycle 20 the BFGS method is 10^{-4} au above the converged energy, whereas both MS algorithms are greater than 10^{-4} au from energy convergence after 50 energy evaluations

We might summarize our findings as follows. For small molecules, the MS and BFGS methods, as we have described them, have similar behavior For all the larger molecules we have examined, those reported here and very many others, the BFGS method is superior in locating minima. Both the MS and BFGS methods seem equally able to reduce the Cartesian gradients

Table 3
BFGS optimization of NH_3 — Hessian eigenvalues (au)

1	1.000D + 00	1.000D + 00	1.000D + 00	1.000D + 00	1.000D + 00	1 000D + 00
2	2.723D + 00	1.000D + 00	1.000D + 00	1.000D + 00	1 000D + 00	9.784D − 01
3	3.719D + 00	1.755D + 00	1.000D + 00	1.000D + 00	1.000D + 00	6 725D − 01
4	2 836D + 00	2 378D + 00	1 032D + 00	1.000D + 00	1.000D + 00	2.641D − 01
5	2.785D + 00	2.377D + 00	1.527D + 00	1.000D + 00	1 000D + 00	1 618D − 01
6	2.577D + 00	2 364D + 00	1.514D + 00	1 000D + 00	9.994D − 01	1.541D − 01
7	2.423D + 00	2.265D + 00	1 476D + 00	1.000D + 00	9.952D − 01	1 627D − 01
9	2 434D + 00	2.347D + 00	1 472D + 00	1 000D + 00	9 254D − 01	1 586D − 01
exact	2 2744D + 00	2 2744D + 00	1 4519D + 00	1 5998D − 01	1 5430D − 01	1 5430D − 01

Evaluation number		Energy	$|g|$	$\max|g_i|$	α
E	G				
1	1	−12 499372	0 358164	0 198326	0 00000D + 00
2	2	−12 521284	0 050917	0.031968	4 00000D − 01
3	3	−12.522072	0 028891	0 018206	1.00000D + 00
4	4	−12 522491	0 019688	0.015225	1 00000D + 00
5	5	−12 523027	0.016294	0 010447	1 00000D + 00
6	6	−12 523199	0 007948	0.005281	1 00000D + 00
7	7	−12 523215	0.000941	0 000621	1 00000D + 00
8	8	−12 523216	0 000101	0 000059	1 00000D + 00

Table 4
A comparison of MS and BFGS update methods for $H_2CCHCHO$ [a]

Cycle	MS1		MS2		BFGS [b]	
	ΔE	$|g|$	ΔE	$|g|$	ΔE	$|g|$
1	0 0123	0 195	0.0123	0 195	0 0123	0 195
5	2.6−3	0 058	2 8−3	0 029	5 0−3	0.028
10	1 3−3	0 012	1.3−3	9.9−3	2 8−?	0 013
20	5 3−4	0.024	4 6−4	6.6−3	1 0−4	4 0−3
30	2 7−4	1 7−3	3 7−4	0.028	1.0−?	2 6−5
40	1.8−4	1.2−3	1.9−4	3.4−3		
50	1 3−4	0.011	1 4−4	2 4−3		

[a] ΔE in au from the energy value, $|g| = (g|g)^{1/2}$ in au/bohr
[b] In 32 cycles the BFGS method has reduced $|g|$ to 2×10^{-6} au/bohr

to below 0.01 au/bohr, but the MS method often has erratic behavior for larger systems after this point The BFGS method *nearly* always reduces the gradient norm in a monotonic fashion, whereas this is seldom the case for the MS method Neither update method produces a reliable Hessian matrix when the maximum gradient has been reduced to 10^{-4} au/bohr, a value at which the geometries are usually reliable. Should a detailed knowledge of the second derivatives be required, then these will have to be evaluated analytically or through numerical differencing of the first derivatives.

Acknowledgement

References

[1] B A Murtagh and R.W H Sargent, Comp. J 13 (1970) 185

[2] I.W McIver Jr. and A Komornicki, J. Am Chem. Soc 94 (1972) 2625

[3] C G. Broyden, J Inst Maths Appl 6 (1970) 76.

[4] R Fletcher, Comp J 13 (1970) 317.

[5] D Goldfarb, Math. Comp 24 (1970) 23.

[6] D F. Shanno, Math Comp. 24 (1970) 647

[7] R Fletcher, in· Practical methods of optimization, Vol 1 (Wiley, New York, 1980)

[8] J.D. Head and M C. Zerner, in preparation.

[9] J Simons, P. Jørgensen, H. Taylor and J. Ozment, J. Phys Chem 87 (1983) 2745

[10] D Spangler, I H. Williams and G M Maggiora, J Comp. Chem 4 (1983) 524.

[11] E B Wilson, J C Decius and P C. Cross, Molecular vibrations (McGraw-Hill, New York, 1955)