The Quasi-Cauchy Relation and Diagonal Updating

M. Zhu*, J.L. Nazareth† and H. Wolkowicz‡

December, 1997

Abstract

The quasi-Cauchy (QC) relation is the weak-secant or weak-quasi-Newton relation of Dennis and Wolkowicz [3] with the added restriction that full matrices are replaced by diagonal matrices. The latter are especially appropriate for scaling a Cauchy (steepest-descent) algorithm, hence our choice of terminology.

In this article, we explore the QC relation and develop variational techniques for updating diagonal matrices that satisfy it. Numerical results are also given to illustrate the use of such updates within a Cauchy algorithm.

Keywords: Weak-secant, Quasi-Cauchy, diagonal updating, Cauchy algorithm, steepest-descent.

1 Introduction

We consider the problem of finding a local minimum of a smooth, unconstrained nonlinear function, namely,

$$\text{minimize}_{x \in R^n} \quad f(x). \tag{1}$$

For a background overview of Newton and Cauchy-type algorithms for solving (1), see Dennis and Schnabel [2] or the recent landmark book of Bertsekas [1].

*Department of Pure and Applied Mathematics, Washington State University, Pullman, WA 99164-3113. E-mail: zhu@delta.math.wsu.edu

†As above. E-mail: nazareth@amath.washington.edu

‡Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario, Canada. E-mail: hwolkowi@orion.math.uwaterloo.ca

In the latter reference, we find the following important observation ([1], p. 67):

> Generally, there is a tendency to think that difficult problems should be addressed with sophisticated methods, such as Newton-like methods. This is often true, particularly for problems with nonsingular local minima that are poorly conditioned. However, it is important to realize that *often the reverse is true*, namely that for problems with "difficult" cost functions and singular local minima, it is best to use simple methods such as (perhaps diagonally scaled) steepest descent with simple stepsize rules such as a constant or a diminishing stepsize. The reason is that methods that use sophisticated descent directions and stepsize rules often rely on assumptions that are likely to be violated on difficult problems.

Our investigation here is very much in the spirit of these remarks. In particular, we seek effective ways to diagonally scale an algorithm of Cauchy type.

For purposes of discussion, it is useful to identify a hierarchy of relations that can be employed within Newton and Cauchy algorithms as follows:

- *Secant* or *Quasi-Newton* (QN): $M_+ s = y$ where the n-dimensional vectors $s = x_+ - x$ and $y = g_+ - g$ denote the step and gradient change corresponding to two different points x and x_+ and their associated gradients g and g_+. M_+ a full $n \times n$ matrix that approximates the Hessian. This notation is used henceforth. Both s and y are available to the associated QN algorithm and it requires $O(n^2)$ storage for the matrix M_+.

- *Weak-Secant*: $s^T M_+ s = s^T y$. This was introduced and studied by Dennis and Wolkowicz [3]. Again the resulting QN algorithm uses s and y explicitly and requires $O(n^2)$ storage.

- *Quasi-Cauchy* (QC): $s^T D_+ s = s^T y$ where D_+ is a *diagonal* matrix, i.e., the QC relation is the weak secant with matrices restricted to be diagonal and s and y are available. The associated Cauchy algorithm requires only $O(n)$ storage.

- *Weak-Quasi-Cauchy*: $s^T D_+ s = b$ where D_+ is a diagonal matrix and $b \equiv s^T g_+ - s^T g = s^T y$ is obtained by directional derivative differences along s, i.e., the weak QC relation is the QC relation further weakened so that gradient vectors (hence the vector y) are not explicitly used. The notions of QC relations and diagonal updating were originally introduced in this setting in [12], [13]. The associated QC algorithm requires $O(n)$ storage and, in addition, only requires approximations to gradients (quasi-gradients).

We will discuss the general idea of diagonal updating subject to the QC relation and give numerical results for an implementation of a Cauchy algorithm that employs such diagonal scaling matrices. A more complete theory of diagonal updating, including its application to limited-memory BFGS algorithms and further numerical results, can be found in [16], [17].

2 Diagonal Updating

Suppose $D > 0$ is a positive definite diagonal matrix and D_+ is the updated version of D which is also diagonal. We require that the updated D_+ satisfy the QC relation and that the deviation between D and D_+ is minimized under some variational principle. We would like the latter to preserve positive definiteness in a natural way, i.e. we seek well-posed *metric* problems such that the solution D_+, through the diagonal updating, incorporates available curvature information from the step and gradient changes as well as that contained in D. As noted earlier, a diagonal matrix simply needs the same computer storage as a vector so an algorithm with $O(n)$ storage will be maintained. We only consider Cauchy algorithms here, but it is clear that diagonal updating will have wide application to CG and limited memory QN algorithms as well.

We now focus on two basic forms of the diagonal updates.

2.1 Updating D

Consider the variational problem:

$$(P): \text{ minimize } \ ||D_+ - D||_F$$

$$\text{s.t. } \ s^T D_+ s = s^T y$$

where $s \neq 0$, $s^T y > 0$ and $D > 0$. Let

$$D_+ = D + \Lambda, \quad a = s^T D s, \quad b = s^T y. \tag{2}$$

Then the variational problem can be stated as

$$(P): \text{ minimize } \ ||\Lambda||_F$$

$$\text{s.t. } s^T \Lambda s = b - a.$$

In (P), the objective is strictly convex and the feasible set is convex. Therefore, there exists a unique solution to (P). Its Lagrangian function is

$$L(\Lambda, \mu) = \frac{1}{2}\mathrm{tr}(\Lambda^2) + \mu(s^T \Lambda s + a - b)$$

where μ is the Lagrange multiplier associated with the constraint and tr denotes the trace operator. Differentiating with respect to Λ via the matrix calculus [6] or differentiating with respect to the diagonal elements, setting the result to zero and invoking the constraint $s^T \Lambda s = b - a$, we have

$$\Lambda = \frac{b-a}{\mathrm{tr}(E^2)} E, \quad E = \mathrm{diag}\,[s_1^2,\, s_2^2,\, \dots,\, s_n^2\,] \tag{3}$$

where s_i is the i'th element of s. When $b < a$, note that the resulting D_+ is not necessarily positive definite. For algorithmic purposes, a safeguard is needed to ensure $D_+ > 0$. This can be easily achieved by checking the condition

$$\forall i, \quad d_i + \frac{(b-a)s_i^2}{\mathrm{tr}(E^2)} > 0 \tag{4}$$

where d_i is the i'th diagonal element of D. When the above is violated, we can retain the previous diagonal matrix by setting $D_+ = D$ or use some simple scheme to generate D_+ such that $D_+ > 0$. An example is to switch to the basic Oren-Luenberger scaling matrix (used in the L-BFGS algorithm), namely,

$$D_+ = (s^T y / s^T s) I$$

where I is the identity matrix. It is useful to note that this is precisely the matrix that would be obtained from the QC relation with the further restriction that the diagonal matrix is a scalar multiple of the identity matrix, i.e., instead of a general diagonal matrix one uses a matrix whose elements on the diagonal are equal.

An algorithm incorporating these details will be considered in the section on numerical results later in this paper.

2.2 Updating $D^{1/2}$

A more efficient way of preserving positive definiteness through diagonal updating is to update the Cholesky[1] factor $D^{1/2}$ to the corresponding $D_+^{1/2}$

[1]'Square-root' would be a more precise choice of terminology, but we use 'Cholesky' to retain the connection with the updating of QN triangular factors of full matrices.

with

$$D_+^{1/2} = D^{1/2} + \Omega$$

and

$$(FP): \text{ minimize } \ ||\Omega||_F$$
$$\text{s.t. } s^T(D^{1/2} + \Omega)^2 s = s^T y > 0.$$

The foregoing variational problem is well-posed, being defined over the closed set of matrices for which the corresponding D_+ is positive-semidefinite. Further, analogously to the full matrix case in standard QN updating, it always has a viable solution for which D_+ is positive definite. This is stated in the following theorem:

Theorem 2.2.1 *Let $D > 0$ and $s \neq 0$, a, b, E are defined in (2) and (3). There is a unique global solution Ω of (FP) which is given by*

$$\Omega = \begin{cases} 0 & \text{if } b = a \\ -\mu^* E(I + \mu^* E)^{-1} D^{1/2} & \text{if } b \neq a \end{cases} \tag{5}$$

where μ^* is the largest solution of the nonlinear equation $F(\mu) = b$ for which

$$F(\mu) \stackrel{\text{def}}{=} s^T(D(I + \mu E)^{-2})s = \sum_{\{i: s_i \neq 0\}} \frac{d_i s_i^2}{(1 + \mu s_i^2)^2} \tag{6}$$

Proof. In the process of the proof we will see every expression above is well defined. First, by some simple transformations, problem (FP) is equivalent to

$$(FP): \text{ minimize } \ ||w||_2^2 = w^T w$$
$$\text{s.t. } w^T E w + 2 w^T E r = b - a$$

where

$$r = [d_1^{1/2},\ d_2^{1/2},\ \ldots,\ d_n^{1/2}]^T$$

When $b = a$, the global optimal solution is obviously $w = 0$, and hence $\Omega = 0$, which implies that $D_+ = D$ is positive definite. In the following discussion we assume that $b \neq a$. Problem (FP) has a strictly convex objective with the Hessian E of the constraint being positive semi-definite. By a theorem in [8] concerning a quadratic objective with a quadratic constraint, (FP) has a global solution. Differentiating its Lagrangian

$$L(w,\ \mu) = w^T w + \mu(w^T E w + 2 w^T E r + a - b)$$

with respect to w, where μ is the Lagrangian multiplier, and setting the result to zero, we have

$$w = -\frac{\mu s_i^2 d_i^{1/2}}{(1+\mu s_i^2)}, \quad i = 1, ..., n$$

Substituting these quantities into the constraint equation, we obtain

$$\begin{aligned} F(\mu) &\stackrel{\text{def}}{=} s^T(D(I+\mu E)^{-2})s \\ &= \sum_{i=1}^{n} \frac{d_i s_i^2}{(1+\mu s_i^2)^2} \\ &= \sum_{\{i:s_i\neq 0\}} \frac{d_i}{s_i^2(\mu + (1/s_i^2))^2} \\ &= b \end{aligned}$$

Note that $F(\mu)$ has poles at $(-1/s_i^2)$, $i = 1, ..., n$. Let

$$j = \arg\max_{\{i:s_i\neq 0\}}(-\frac{1}{s_i^2}).$$

The derivative of $F(\mu)$ is

$$\frac{dF(\mu)}{d\mu} = -2 \sum_{\{i:s_i\neq 0\}} \frac{r_i^2}{s_i^2(\mu + (1/s_i^2))^3} < 0$$

on the interval

$$(-\frac{1}{s_j^2}, +\infty)$$

so $F(\mu)$ is strictly decreasing in the above interval from $+\infty$ to 0. Noting that $b > 0$, we see that there is a unique solution μ^* within this interval such that $F(\mu^*) = b$. Though the behavior of $F(\mu)$ is complex in the entire domain, solutions for $F(\mu) = b$ except μ^* are of no interest (note that μ^* is the largest solution). This is because a necessary condition [8] of the solution of (FP) requires the Hessian of the Lagrangian, namely, $I + \mu E$, to be positive semi-definite. This is equivalent to

$$1 + \mu s_i^2 \geq 0, \quad i = 1, ..., n,$$

and clearly μ^* is the unique solution of $F(\mu) = b$ satisfying the above inequalities. A key observation is that $I + \mu^* E$ is positive definite, and thus

μ^* is the unique global minimizer for (FP). Returning to the relationship of w and μ, we see that

$$w^* = -\mu^* E(I + \mu^* E)^{-1} D^{-1/2}$$

is the unique solution of (FP). Note also that $\forall i = 1, ..., n$,

$$d_i^{1/2} - \frac{\mu^* s_i^2 d_i^{1/2}}{(1 + \mu^* s_i^2)} = \frac{1}{1 + \mu^* s_i^2} d_i^{1/2} \neq 0$$

so D_+ is positive definite. This completes the proof. ■

The following is a direct result of the theorem.

Corollary 2.2.1 *The solution D_+ through the diagonal updating problem (FP) is positive definite and unique which is given by:*

$$D_+ = \begin{cases} D & \textit{if } b = a \\ (I + \mu^* E)^{-2} D & \textit{if } b \neq a \end{cases} \tag{7}$$

2.3 Discussion

Suppose n is not large and that evaluating a function/gradient is relatively expensive (a common assumption in nonlinear optimization). Then the cost of solving the nonlinear equation $F(\mu) = b$, which we call the QC subproblem henceforth, is essentially trivial, even when it is performed by a crude unidimensional algorithm, for example, bisection. If greater efficiency is needed, it is useful to exploit a connection[2] between the QC subproblem and a scaled trust-region subproblem derived from a reformulation of (FP) as follows:

$$\text{minimize } \; ||D_+^{1/2} - D^{1/2}||_F$$
$$\text{s.t. } \; s^T D_+ s = b > 0.$$

Then using the earlier definitions

$$E = \text{diag}\,[s_1^2,\ s_2^2,\ \dots, s_n^2\ ],$$

$$r = [d_1^{1/2},\ d_2^{1/2},\ ...,\ d_n^{1/2}\ ]^T,$$

[2]This connection is particularly ironic, because the QC method developed in this article is quintessentially *metric-based*, whereas trust-region techniques are the fundamental building blocks of *model-based* approaches—for terminology see Nazareth [11].

and defining the vector z to be the diagonal elements of $D_+^{1/2}$, we can reexpress the foregoing variational problem as follows:

$$\text{minimize} \quad - r^T z + \frac{1}{2} z^T z$$

$$\text{s.t.} \;\; z^T E z = b$$

where $b > 0$. When E is nonsingular (hence positive definite) and the equality in the constraint is replaced by a $\leq$ inequality, one obtains a standard trust-region subproblem in the metric defined by E. The QC subproblem can be viewed as a simple but nonstandard trust region problem[3]. Thus many of the techniques used to solve trust-region subproblems—see, in particular, Rendl and Wolkowicz [15]—can be suitably adapted to solving the QC subproblem more efficiently. Our purpose in the present article is to explore the QC approach at a root level and further refinements will be considered in a subsequent paper including comparison with recent *non*-monotonic Cauchy-based algorithms, see Raydan [14].

3 Numerical Results

In this section we give some numerical results on the application of diagonal updating to the Cauchy algorithm. Diagonal updating can be used as a dynamical scaling at each iteration to the steepest descent direction in the Cauchy algorithm. The Cauchy direction is ideal when the contours of the objective f to be minimized are hyperspheres. For a general function which is not quadratic, a preconditioning can be used to make the transformed contours closer to hyperspheres such that the the efficiency of the Cauchy direction in the transformed space is enhanced, see [11]. The diagonal updating is a nonfixed preconditioning which includes the updated curvature information, and its hereditary positive definiteness is naturally maintained when the Cholesky factor is updated as shown in the previous section. An

[3] *Simple* because only diagonal matrices are involved so issues associated with cost of matrix inversions or factorizations of a more general quadratic objective do not arise. Also, because all components of r are positive and the eigenvectors associated with the Hessian of the objective (or any diagonal rescaling of it) are along the coordinate axes, which leads to theoretical and algorithmic simplifications. In particular, r has a nonzero component in the eigenspace associated with the smallest eigenvalue. *Nonstandard* because $z = r$ is not an acceptable solution of the QC problem when $r^T E r < b$ as in the usual inequality constrained trust-region problem. Also, because E can be singular, in which case the corresponding components of z are separable and can be set to the components of r.

Number	Problem Name
1	Helical valley function
2	Biggs exp6 function
3	Gaussian function
4	Powell badly scaled function
5	Box 3-dimensional function
6	Variably dimensioned function
7	Watson function
8	Penalty function I
9	Penalty function II
10	Brown badly scaled function
11	Brown and Dennis function
12	Gulf research and development function
13	Trigonometric function
14	Extended Rosenbrock function
15	Extended Powell function
16	Beale function
17	Wood function
18	Chebyquad function

Table 1: Test Problems

expectation that the Cauchy method will be significantly accelerated using diagonal updating is supported by our numerical results.

Our source code is written in Fortran 90, with double precision algorithmic, running on an ULTRIX DEC Alpha workstation. The numerical experiment is done within the MINPACK-1 testing environment [10]. Test functions are the standard unconstrained problems collected in [7], which we identify by the numbering in Table 1.

We employ a line search (rather than the more simple stepsize choices mentioned in the quotation at the beginning of this paper) and use a routine of Moré and Thuente [9] based on cubic interpolation, which satisfies the Wolfe conditions:

$$f(x_+) \leq f(x) + \alpha\lambda g^T d \tag{8}$$

$$g(x_+)^T d \geq \beta g^T d \tag{9}$$

where the line search parameters are chosen as [4]: $\alpha = 10^{-4}, \beta = 0.9$. The stopping criterion is [4]:

$$||g(x)|| \leq 10^{-5} max\{1.0, ||x||\} \tag{10}$$

The methods tested include:

1. *Standard Cauchy* algorithm of the simple form $d = -g$ at the k'th iteration.

2. *Cauchy with Oren-Luenberger Scaling*: this scales the search direction with the well-known Oren-Luenberger Scaling [5]:

$$d = -\frac{y^T s}{y^T y} g$$

   for all the iterations after the first step where the initial steepest descent search is employed.

3. *Cauchy-DU*: Cauchy algorithm with diagonal updating, i.e., at the current iterate the search direction d is scaled from the steepest descent as:

$$d = -U_+ g$$

   where U_+ is updated from $U = D^{-1}$, which corresponds to complementary diagonal updating:

$$(CP): \text{ minimize } \; ||U_+ - U||_F$$

$$\text{s.t. } \; y^T U_+ y = y^T s$$

   For the details about the complementarity on (P) and (CP), see [16]. The updated diagonal matrix is given by

$$U_+ = U + \Gamma = U + \frac{b-c}{\text{tr}(G^2)} G$$

   where

$$c = y^T U y, \;\; G = \text{diag}\,[\; y_1^2, \; y_2^2, \; ..., \; y_n^2 \;]$$

   with the safeguarding policy as follows: the above updating is used only when the condition

$$\forall i, \;\; u_i + \frac{(b-c)y_i^2}{\text{tr}(G^2)} > 0$$

is satisfied (u_i are the diagonal elements of U). Otherwise the constant diagonal matrix as the basic matrix in the L-BFGS algorithm is used, i.e.

$$U_+ = (y^T s/y^T y)I \tag{11}$$

For algorithmic consideration of L-BFGS, see [4] and [16].

4. *Cauchy-Cholesky*: Cauchy algorithm with the diagonal updating for the Cholesky factor $U^{1/2}$, where again considering the complementary problem we have

$$U_+ = \begin{cases} U & \text{if } b = c \\ (I + \nu^* G)^{-2} U & \text{if } b \neq c \end{cases} \tag{12}$$

where ν^* is the largest solution for $H(\nu) = b$ for which

$$H(\nu) \stackrel{\text{def}}{=} y^T (U(I + \nu G)^{-2}) y = \sum_{\{i : y_i \neq 0\}} \frac{u_i y_i^2}{(1 + \nu y_i^2)^2}$$

In our numerical implementation, ν^* is obtained by either a Newton algorithm for a unidimensional function, or a simple bisectional searching within the interval from the largest pole of the function $H(\nu)$ to some large number in the axis such that the initial bisection condition of the endpoints is satisfied. Note that $H(0) = c$, thus if $b > c$, then the solution $\nu^* < 0$; if $b < c$, then $\nu^* > 0$. And hence the interval for the bisection is actually reduced with one endpoint being 0 in each case. Also a Newton step for searching for the solution of $H(\nu) = b$ always starts from zero. (Note that more efficient reformulations and techniques for solving the QC subproblem are possible as discussed in Section 2.3.)

The numerical comparative results are given in the following tables. In all the tables we give the $nitr/nfg$ as the number of iterations and effective calls for function and gradient evaluation. The symbol $*$ in the table indicates that the method takes too many iterations and is regarded as having failed to converge. The first and second columns in the tables are the numbers standing for the test problems and the problem dimensions, respectively. The remaining columns are the results for the corresponding methods.

From the above results we see the Cauchy algorithms using diagonal updating are much faster than the standard Cauchy. And in most problems the diagonal updating for the Cholesky factor performs better than the ad hoc

Prob	Dim	Cauchy	Cauchy-OL	Cauchy-DU	Cauchy-Cholesky
1	3	2552/5229	431/756	378/708	370/688
2	6	24041/45488	2221/4353	2977/5762	1165/2120
3	3	2/4	2/6	2/6	2/6
4	2	*	*	238/1649	238/1649
5	3	32535/65075	225/428	474/914	165/300
6	6	446/1001	574/877	120/254	157/274
6	8	981/2318	269/415	184/332	229/427
7	2	14/35	15/20	22/26	15/20
8	4	46282/46295	491/1386	783/2327	491/1386
9	4	63/128	40/61	84/93	49/66
10	2	*	147/998	*	147/998
11	4	*	126/892	121/617	198/387
12	3	*	988/2506	1776/4530	*
13	4	76/93	35/46	40/53	67/85
13	8	134/169	109/156	75/99	80/120
14	2	1109/2248	242/558	408/995	289/701
15	4	70638/159377	2853/5081	1040/2157	428/827
16	2	188/377	315/471	186/276	104/167
17	4	2879/5795	1755/2347	2022/3714	525/1003
18	4	11/25	16/21	18/22	16/20
18	8	118/253	82/128	64/94	67/98

Table 2: Numerical Results for Diagonal Updating

diagonal updating strategy with the safeguard policy for positive definiteness. The DU-Cauchy is competitive with the Cholesky form because the former can be implemented very simply whereas the latter incurs an overhead for computing the optimal ν for a highly nonlinear one-dimensional equation. But from a wider perspective, the Cholesky form of diagonal updating is very successful in accelerating the Cauchy algorithm and the expense of solving ν is a relatively minor portion of the total algorithm itself. The Cauchy-OL is competitive due to its simple formulation, and indeed there are some cases in the table for which it requires fewer iterations and function/gradient calls than diagonal updating. But it is clear that globally the Cauchy-Cholesky is best. In comparison, the results for Oren-Luenberger scaling fluctuate greatly. This variability of performance has already been observed in the literature including [5] where even for the simple BFGS algorithm, the Oren-Luenberger scaling for the Hessian matrix, namely $(s^T y/s^T s)I$, does not consistently reduce the iteration and function/gradient calls vis-a-vis the pure BFGS method. Hence, the above results show that diagonal updating could be a more stable scaling in practice.

4 Conclusion

Diagonal updating is a fascinating theory whose appeal arises from its simplicity, elegant solutions and the similarity of the variational metrics employed to those of Quasi-Newton methods, e.g., BFGS, SR1 and LPD. A thorough exploration of both theoretical and practical aspects is ongoing, and further results, in particular, the use of diagonal updating within the L-BFGS algorithm, can be found in [16] and [17].

References

[1] Bertsekas, D.P. (1995), *Nonlinear Programming*, Athena Scientific, Belmont, Massachusetts.

[2] Dennis, J.E. and Schnabel, R.B. (1983), *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, New Jersey.

[3] Dennis, J.E. and Wolkowicz, H. (1990), *Sizing and Least Change Secant Updates*, CORR Report 90-02, Department of Combinatorics and Optimization, University of Waterloo, Ontario, Canada.

[4] Liu, D.C, Nocedal, J., (1989), *On the limited memory method for large scale optimization*, Mathematical Programming B, 45, 3, 503-528.

[5] Luenberger, D.G., (1984), *Linear and Nonlinear Programming*, Addison-Wesley (second edition), 1984.

[6] Magnus, R.J. (1988), *Matrix differential calculus with applications in statistics and econometrics*, Wiley, Chichester.

[7] Moré, J.J., Garbow, B.S. and Hillstrom, K.E. (1981), *Testing unconstrained optimization software*, ACM Transactions on Mathematical Software, 7, 17-41.

[8] Moré, J.J. (1993), *Generalizations of the trust region problem*, Optimization Methods and Software, 2, 189-209.

[9] Moré, J.J. and Thuente. D.J. (1994), *Line search algorithms with guaranteed sufficient decrease*, ACM Transactions on Mathematical Software, 20, 1994, June Issue.

[10] Moré, J.J. and Averick, B. A. (1994), *User Guide for the MINPACK-2 Test Problem Collection,* Technical Memorandum ANL/MCS-TM-157.

[11] Nazareth, J.L. (1994), *The Newton-Cauchy Framework: A Unified Approach to Unconstrained Nonlinear Minimization*, LNCS 769, Springer-Verlag, Berlin.

[12] Nazareth, J.L. (1995), *If quasi-Newton then why not quasi-Cauchy ?*, SIAG/OPT Views-and-News, 6, 11-14.

[13] Nazareth, J.L. (1995), *The Quasi-Cauchy method: a stepping stone to derivative-free algorithms*, TR 95-3, Department of Pure and Applied Mathematics, Washington State University.

[14] Raydan, M. (1993), "On the Barzalai and Borwein choice of steplength for the gradient method", *IMA J. Numerical Analysis*, 13, 321-326.

[15] Rendl, F. and H. Wolkowicz (1997), *A semidefinite framework for trust region subproblems with applications to large scale minimization*, Mathematical Programming, 77, 273-300.

[16] Zhu, M. (1997), *Limited memory BFGS algorithms with diagonal updating*, Project of M.S. in Computer Science, School of Electrical Engineering and Computer Science, Washington State University.

[17] Zhu, M. (1997), *Techniques for Large-Scale Nonlinear Optimization: Principles and Practice*, Ph.D dissertation, Department of Pure and Applied Mathematics, Washington State University.