Ladislav Lukšan; Jan Vlček
A hybrid method for nonlinear least squares that uses quasi-Newton updates applied to an approximation of the Jacobian matrix


Persistent URL: http://dml.cz/dmlcz/703066

Terms of use:
© Institute of Mathematics CAS, 2019

Institute of Mathematics of the Czech Academy of Sciences provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these Terms of use.

This document has been digitized, optimized for electronic delivery and stamped with digital signature within the project DML-CZ: The Czech Digital Mathematics Library http://dml.cz
A HYBRID METHOD FOR NONLINEAR LEAST SQUARES
THAT USES QUASI-NEWTON UPDATES APPLIED
TO AN APPROXIMATION OF THE JACOBIAN MATRIX

Ladislav Lukšan, Jan Vlček
Institute of Computer Science, The Czech Academy of Sciences
Pod Vodárenskou věží 2, 182 07 Prague 8, Czech Republic
luksan@cs.cas.cz, vlcek@cs.cas.cz

Abstract: In this contribution, we propose a new hybrid method for minimization of nonlinear least squares. This method is based on quasi-Newton updates, applied to an approximation $A$ of the Jacobian matrix $J$, such that $A^T f = J^T f$. This property allows us to solve a linear least squares problem, minimizing $\|Ad + f\|$ instead of solving the normal equation $A^T Ad + J^T f = 0$, where $d \in \mathbb{R}^n$ is the required direction vector. Computational experiments confirm the efficiency of the new method.

Keywords: Nonlinear least squares, hybrid methods, trust-region methods, quasi-Newton methods, numerical algorithms, numerical experiments.

MSC: 65K10, 65F30

1. Introduction
Consider the objective function

$$ F(x) = \frac{1}{2} f^T(x) f(x) = \frac{1}{2} \sum_{k=1}^{m} f_k^2(x), \quad (1) $$

where $f : \mathbb{R}^n \to \mathbb{R}^m$ is a twice continuously differentiable mapping with elements $f_k(x)$, $1 \leq k \leq m$. Let $J(x)$ be its Jacobian matrix with elements $J_{kl}(x) = \partial f_k(x)/\partial x_l$, where $1 \leq k \leq m$ and $1 \leq l \leq n$. Then the gradient and the Hessian matrix of function (1) have the form

$$ g(x) = J^T(x) f(x) = \sum_{k=1}^{m} f_k(x) g_k(x), \quad (2) $$

$$ G(x) = J^T(x) J(x) + C(x) = \sum_{k=1}^{m} g_k(x) g_k^T(x) + \sum_{k=1}^{m} f_k(x) G_k(x), \quad (3) $$

DOI: 10.21136/panm.2018.11

99
where \( g_k(x) \) and \( G_k(x) \) are gradients and Hessian matrices of functions \( f_k(x) \), \( 1 \leq k \leq m \). The most known methods for minimization of the objective function (1) are trust-region realizations of the Gauss-Newton method, which are iterative and their iterations have the form

\[
x_{i+1} = x_i, \quad \Delta_{i+1} < \Delta_i \quad \text{if} \quad \frac{F(x_i + d_i) - F(x_i)}{Q_i(d_i)} < \rho,
\]

\[
x_{i+1} = x_i + d_i, \quad \Delta_{i+1} \geq \Delta_i \quad \text{if} \quad \frac{F(x_i + d_i) - F(x_i)}{Q_i(d_i)} \geq \rho,
\]

where \( 0 < \rho < 1 \), \( Q_i(d) = g(x_i)^T d + (1/2)d^T B_i d \), \( B_i = J(x_i) J(x_i) \) and \( d_i \) is an approximate minimum of the quadratic function \( Q_i(d) \) on the trust region defined by constraint \( \|d\| \leq \Delta_i \) [2], [8]. Let \( x^* \in \mathbb{R}^n \) be a minimum of function (1). The Gauss-Newton method works well if \( F(x^*) \) is small (if \( F(x^*) = 0 \), the rate of convergence is superlinear), but the convergence can be slow if \( F(x^*) \) is large. Thus hybrid methods, which are combinations of the Gauss-Newton method and variable metric methods, are advantageously used. In the subsequent text, we use the notation \( F_i = F(x_i), g_i = g(x_i), G_i = G(x_i) \), etc, and \( F^* = F(x^*), g^* = g(x^*), G^* = G(x^*) \), etc. Sometimes index \( i \) is omitted and index \( i+1 \) is replaced by the symbol \( + \). More details concerning methods described in this contribution can be found in [6].

### 2. Hybrid methods

Hybrid methods are based on the fact that \( (F_i - F_{i+1})/F_i \rightarrow 1 \), if \( F_i \rightarrow F^* = 0 \) \( Q \)-superlinearly, and \( (F_i - F_{i+1})/F_i \rightarrow 0 \), if \( F_i \rightarrow F^* > 0 \). This fact forms the basis for a simple hybrid method in [1]: Let \( B_1 = J_i^T J_i \). If \( (F_i - F_{i+1})/F_i \geq \underline{\varrho} \), we set \( B_{i+1} = J_{i+1}^T J_{i+1} \). If \( (F_i - F_{i+1})/F_i < \underline{\varrho} \), we set

\[
B_{i+1} = \frac{1}{\gamma_i} \left( B_i + [y_i, B_is_i] M_i^B [y_i, B_is_i]^T \right),
\]

where \( s_i = x_{i+1} - x_i, \quad y_i = g_{i+1} - g_i = J_i^T f_{i+1} - J_i^T f_i, \quad \gamma_i > 0 \) and the matrix \( M_i^B \in \mathbb{R}^{2 \times 2} \) is chosen in such a way that the quasi-Newton condition \( B_{i+1}s_i = y_i \) is satisfied [7]. This simple hybrid method switches between the Gauss-Newton method and a selected variable metric method (defined by matrix \( M_i^B \)). More complicated hybrid methods are based on structured variable metric updates [4]: Let \( C_1 = 0 \) and \( B_i = J_i^T J_i + C_i \). If \( (F_i - F_{i+1})/F_i \geq \underline{\varrho} \), we set \( C_{i+1} = 0 \). If \( (F_i - F_{i+1})/F_i < \underline{\varrho} \), we set

\[
C_{i+1} = \frac{1}{\gamma_i} \left( C_i + [z_i, C_is_i] M_i^C [z_i, C_is_i]^T \right),
\]

where \( s_i = x_{i+1} - x_i, \quad z_i = J_{i+1}^T f_{i+1} - J_i^T f_i, \quad \gamma_i > 0 \) and the matrix \( M_i^C \in \mathbb{R}^{2 \times 2} \) is chosen in such a way that the quasi-Newton condition \( C_{i+1}s_i = z_i \) is satisfied, see [7].

If matrix \( B_i \) is ill-conditioned, then a more advantageous way is to use a full rank approximation \( A_i \) of the Jacobian matrix \( J_i \) and replace the solution of the normal
equation \( d_i = -B^{-1}_i g_i \), where \( B_i = A_i^T A_i \), by the solution of the linear least-squares problem \( d_i = -A_i^T f_i \). This approach is used in [13], where matrix \( A_i \) is expressed as a sum \( A_i = J_i + L_i \) and matrix \( L_i \) is updated to satisfy the quasi-Newton condition \((J_{i+1} + L_{i+1})^T (J_{i+1} + L_{i+1}) s_i = y_i \). This approach is not quite rigorous, since usually \( A_i^T f_i \neq B_i^{-1} g_i \). The equality \( A_i^T f_i = B_i^{-1} g_i \) is satisfied only if \( A_i^T f_i = g_i = J_i^T f_i \).

For this purpose, the additional condition \( L_{i+1} f_{i+1} = 0 \) was added to the above quasi-Newton condition in [11]. In this contribution, we confine our attention to the simple hybrid method of the form \((6)\). By a variational principle, we derive the update, which satisfies the quasi-Newton condition \( A_{i+1}^T A_{i+1} s_i = y_i \) together with the condition \( A_{i+1}^T f_{i+1} = g_{i+1} = J_{i+1}^T f_{i+1} \).

### 3. New hybrid method

Let \( B = A^T A \), where \( A = J \) if the Gauss-Newton step is accepted, so \( B = J^T J \) holds. To use the variational principle, we write the standard quasi-Newton condition \( B_i s = A_i^T A_i s = y \) in the form

\[
\sqrt{\gamma} A_i s = \tilde{z}, \quad \sqrt{\gamma} A_i^T \tilde{z} = \gamma y, \quad \tilde{z}^T \tilde{z} = \gamma s^T y, \quad (8)
\]

where \( \tilde{z} \in \mathbb{R}^m \) is a free vector parameter. Notice that the last equality, which is a consequence of the first two equalities, is the only restriction on the choice of \( \tilde{z} \).

**Theorem 1.** Let \( W \) be a symmetric positive definite matrix. Then the Frobenius norm \( \|W^{-1/2}(\sqrt{\gamma} A_i - A)^T\|_F \) is minimal on the set of all matrices satisfying quasi-Newton condition \((8)\) if and only if

\[
\sqrt{\gamma} A_i = A^T - \frac{W s}{s^T W s} s^T + \left( \gamma y - z + s^T \tilde{z} \right) \frac{W s}{s^T W s} \frac{\tilde{z}^T}{\tilde{z}^T \tilde{z}}, \quad \tilde{z}^T \tilde{z} = \gamma s^T y, \quad (9)
\]

where \( \tilde{s} = As \) and \( z = A^T \tilde{z} \).

**Proof.** This proof is similar to the proof of Theorem 3.1 proposed in [12]. Denote \( X = \sqrt{\gamma} A_i^T \). Necessity will be proven using the Lagrangian function

\[
L = \frac{1}{2} \left\| W^{-1/2} (X - A^T) \right\|_F^2 + \bar{u}^T (X^T s - \bar{z}) + \bar{v}^T (X \bar{z} - \gamma y) = \sum_{i=1}^{m} \left[ \frac{1}{2} (\xi_i - a_i)^T W^{-1} (\xi_i - a_i) + \bar{u}_i s^T \xi_i + \bar{z}_i v^T \xi_i \right] - \bar{u}^T \bar{z} - \gamma v^T s,
\]

where \( A^T = [a_1, \ldots, a_m] \) and \( X = [\xi_1, \ldots, \xi_m] \). Differentiating the Lagrangian function we obtain

\[
\frac{\partial L}{\partial \xi_i} = W^{-1} (\xi_i - a_i) + \bar{u}_i s + \bar{z}_i v.
\]

Therefore, the conditions for stationarity of the Lagrangian function have the form

\[
W^{-1}(\xi_i - a_i) + \bar{u}_i s + \bar{z}_i v = 0, \quad 1 \leq i \leq m,
\]

or

\[
X - A^T = -W s \bar{u}^T - W v \bar{z}^T.
\]
Using the first condition from (8) we obtain
\[ X^T s = As - s^T W s \bar{u} - v^T W s \tilde{z} = \tilde{z} \quad \Rightarrow \quad \bar{u} = \frac{1}{s^T W s} (As - (1 + v^T W s) \tilde{z}), \]
which after substitution to the previous equality gives
\[ X - A^T = -\frac{W s}{s^T W s} s^T + wz^T, \]
where \( w \in \mathbb{R}^n \) is an unknown vector (determined uniquely by vector \( v \)). Using the second condition from (8) we obtain
\[ X \tilde{z} = A^T \tilde{z} - s^T A^T \tilde{z} \frac{W s}{s^T W s} + \tilde{z} w = \gamma y \quad \Rightarrow \quad w = \frac{1}{\tilde{z}^T \tilde{z}} \left( \gamma y - z + \frac{s^T z}{s^T W s} W s \right), \]
which after substitution to the previous equality (with using relation \( X = \sqrt{\gamma} A^T \)) gives (9). Sufficiency follows from the convexity of the Frobenius norm.

Update (9) contains two vector parameters \( W s/s^T W s \) and \( \tilde{z} \). These parameters should be chosen in such a way to guarantee condition \( A^T f_+ = g_+ \).

**Lemma 1.** Equalities
\[ \sqrt{\gamma} A_+ s = \tilde{z}, \quad \sqrt{\gamma} A_+^T \tilde{z} = \gamma y, \quad A_+^T f_+ = g_+ \tag{10} \]
can be satisfied simultaneously only if
\[ f_+^T f_+ s^T y \geq (s^T g_+)^2. \tag{11} \]

**Proof.** From the first two equalities in (10), the relation \( \tilde{z}^T \tilde{z} = \gamma s^T y \) follows, which determines the norm of vector \( \tilde{z} \). The first and the third equalities imply \( f_+^T \tilde{z} = \sqrt{\gamma} f_+^T A_+ s = \sqrt{\gamma} s^T g_+ \). Since the distance of the hyperplane \( f_+^T \tilde{z} = \sqrt{\gamma} s^T g_+ \) from the origin is equal to \( \sqrt{\gamma} |s^T g_+|/\|f_+\| \), the norm of vector \( \tilde{z} \) cannot be smaller than this number, which together with equality \( \|\tilde{z}\| = \sqrt{\gamma} s^T y \) gives \( \sqrt{\gamma} |s^T g_+|/\|f_+\| \leq \sqrt{\gamma} s^T y \), or \( f_+^T f_+ s^T y \geq (s^T g_+)^2 \).

**Remark 1.** If perfect line search is used, then \( s_i^T g_{i+1} = 0 \) holds in every iteration, so \( s_i^T y_i = s_i^T g_{i+1} - s_i^T g_i = -s_i^T g_i > 0 \), and condition (11) is always satisfied. If the strong Wolfe condition is used (see [10]), then \( |s_i^T g_{i+1}| \leq \varepsilon_2 |s_i^T g_i| \) holds in every iteration, so \( s_i^T y_i = s_i^T g_{i+1} - s_i^T g_i \geq (1 - \varepsilon_2)|s_i^T g_i| \), and condition (11) is satisfied whenever
\[ f_{i+1}^T f_{i+1} \geq \frac{\varepsilon_2^2}{1 - \varepsilon_2} |s_i^T g_i|. \tag{12} \]
If \( x_i \to x^* \) (so \( g_i \to 0 \) and \( s_i \to 0 \)) and \( F(x^*) > 0 \), there exists an index \( k \in \mathbb{N} \) such that condition (12) (and therefore also condition (11)) is satisfied \( \forall i \geq k \). Moreover, in our numerical experiments with Algorithm 1, the condition (11) was always satisfied, if \( F_i - F_{i+1} \leq \frac{1}{2} F_i \) with \( \bar{\varepsilon} = 0.0005 \).

102
Theorem 2. Let vectors $f_+$ and $A$ be linearly independent and assume the inequality (11) holds. If we use vectors $\tilde{\xi} = \sqrt{\gamma}(\lambda_1 f_+ + \lambda_2 A)$, where
\[
\lambda_2^2 = \frac{s^T y f_+^T f_+ - (s^T g_+)^2}{f_+^T s^T A As - (s^T A f_+)^2}, \quad \lambda_1 = \frac{s^T g_+ - \lambda_2 s^T A f_+}{f_+^T f_+},
\]
and
\[
\frac{W s}{s^T W s} = \frac{\gamma s^T y (A^T f_+ - \sqrt{\gamma} g_+) + \sqrt{\gamma} s^T g_+ (\gamma y - A^T \tilde{\xi})}{\gamma s^T y s^T A f_+ - \sqrt{\gamma} s^T g_+ s^T A \tilde{\xi}},
\]
in formula (9), then equalities (10) hold.

Proof. Vector $\tilde{\xi}$ has to satisfy equalities $f_+^T \tilde{\xi} = \sqrt{\gamma} s^T g_+$ and $\tilde{\xi}^T \tilde{\xi} = \gamma s^T y$. Setting $\tilde{\xi} = \sqrt{\gamma}(\lambda_1 f_+ + \lambda_2 A)$, we obtain the system of equations
\[
\begin{align*}
\lambda_1 f_+^T f_+ + \lambda_2 s^T A f_+ &= \sqrt{\gamma} s^T g_+, \\
\lambda_2^2 f_+^T f_+ + 2 \lambda_1 \lambda_2 s^T A f_+ + \lambda_2^2 s^T A A s &= \gamma s^T y
\end{align*}
\]
for unknowns $\lambda_1$ and $\lambda_2$. Since the vectors $f_+$ and $A$ are linearly independent, these equations have the unique solution given by (14). Update (9) satisfies the first two equalities in (10) (Theorem 1). Using the third equality, we obtain
\[
\sqrt{\gamma} g_+ = A^T f_+ - \frac{W s}{s^T W s} s^T A f_+ + \left(\gamma y - A^T \tilde{\xi} + s^T A \tilde{\xi} \frac{W s}{s^T W s}\right) \frac{\tilde{\xi}^T f_+}{\tilde{\xi}^T \tilde{\xi}} = A^T f_+ - \left(s^T A f_+ - s^T A \tilde{\xi} \frac{\sqrt{\gamma} s^T g_+}{\gamma s^T y}\right) w + \left(\gamma y - A^T \tilde{\xi}\right) \frac{\sqrt{\gamma} s^T g_+}{\gamma s^T y},
\]
where $w = W s / s^T W s$. This relation implies that
\[
w = \lambda \left(A^T f_+ - \sqrt{\gamma} g_+ + (\gamma y - A^T \tilde{\xi}) \frac{\sqrt{\gamma} s^T g_+}{\gamma s^T y}\right),
\]
and since $s^T w = s^T W s / s^T W s = 1$, one can write
\[
\lambda \left(\gamma s^T y s^T A f_+ - \sqrt{\gamma} s^T g_+ s^T A \tilde{\xi}\right) = \gamma s^T y.
\]
Substituting this value $\lambda$ into (16), we obtain (15).}

The above considerations are summarized in the following algorithm.

Algorithm 1

Data: Trust-region parameters [8], update parameter $\vartheta = 0.0005$, termination parameters $\varepsilon = 10^{-15}$, $\varepsilon = 10^{-5}$.

Step 1: Initiation. Choose starting point $x_1 \in \mathbb{R}^n$ and initial trust-region radius $\Delta_1 > 0$. Compute $f_1 = f(x_1)$, $J_1 = J(x_1)$, $F_1 = (1/2)f_1^T f_1$, $g_1 = J_1^T f_1$. Set $A_1 = J_1$ and $i = 1$.
Step 2: Termination. If $F_i \leq \varepsilon$ or $\|g_i\| \leq \varepsilon$, then terminate the computation.

Step 3: Direction determination. Determine direction vector $d_i$ using a trust-region strategy (see [8]). Compute $f(x_i + d_i)$, $F(x_i + d_i) = (1/2)f(x_i + d_i)^Tf(x_i + d_i)$. Determine $x_{i+1}$ and $\Delta_{i+1}$ by (4)–(5).

Step 4: Decision. If $x_{i+1} = x_i$, go to Step 2. If $x_{i+1} \neq x_i$, set $f_{i+1} = f(x_i + d_i)$, $F_{i+1} = F(x_i + d_i)$ and compute $J_{i+1} = J(x_i + d_i)$, $g_{i+1} = g(x_i + d_i)$.

Step 5: Update. If $(F_i - F_{i+1})/F_i \geq \vartheta$, set $A_{i+1} = J_{i+1}$. If $(F_i - F_{i+1})/F_i < \vartheta$, compute matrix $A_{i+1}$ by (9) with (13)–(15).

Step 6: Increase $i$ by 1 and go to Step 2.

4. Computational experiments

Methods for nonlinear least-squares were tested by using 80 problems with 200 variables taken from the collection TEST24 contained in the software system for universal functional optimization UFO [9]. Table 1 contains results obtained by the following methods:

<table>
<thead>
<tr>
<th>Method</th>
<th>Rectangular matrix update</th>
<th>QR decomposition update</th>
</tr>
</thead>
<tbody>
<tr>
<td>GN</td>
<td>NIT 3376</td>
<td>NFV 3698</td>
</tr>
<tr>
<td>HN</td>
<td>NIT 2477</td>
<td>NFV 2730</td>
</tr>
<tr>
<td>HS</td>
<td>NIT 2477</td>
<td>NFV 2716</td>
</tr>
<tr>
<td>QN</td>
<td>NIT 5473</td>
<td>NFV 5988</td>
</tr>
<tr>
<td>QB</td>
<td>NIT 6904</td>
<td>NFV 8092</td>
</tr>
</tbody>
</table>

Table 1: TEST24 – 80 problems with 200 variables
The results contained in Table 1 imply several conclusions:

- Hybrid methods \(\text{HN}\) and \(\text{HS}\) are more robust than Gauss-Newton method \(\text{GN}\), since they increase the rate of convergence for large residual problems. The new method seems to be better than structured hybrid method \(\text{HS}\), especially if the efficiency is measured by the computational time.

- Quasi-Newton methods [8], developed originally for solving nonlinear equations, are surprisingly efficient, if they are applied to the QR decomposition of the matrix \(A\), especially if the efficiency is measured by the computational time.

For better understanding, the methods that update rectangular matrix (the first part of Table 1) are also compared by using performance profiles proposed in [5]. In Figure 1, value \(\rho_M(0)\) is the percentage of the test problems for which method \(M\) is the best and value \(\rho_M(\tau)\) for \(\tau\) large enough is the percentage of the problems that method \(M\) can solve. Performance profiles show the relative efficiency and reliability of the methods: the higher is the particular curve, the better is the corresponding method.

Notice that the Gauss-Newton method has a better score for \(\tau = 0\), since 80% of problems used have zero residuals.

**Acknowledgements**

This work was supported by the Institute of Computer Science of the CAS (RVO: 67985807).

**References**


