JOURNAL OF COMPUTATIONAL PHYSICS 22, 534–542 (1976)

Note

On the Convergence of Standard and Damped Least Squares Methods

INTRODUCTION

While the use of least squares minimization is quite commonly used, the Newton–Raphson algorithm often fails to converge or converges very slowly for nonlinear problems. The convergence is known to become poorer with increasing nonlinearity, and an important increase in the quality of the initial estimate of the parameters is needed to reach the solution. In problems where parameters appear in highly nonlinear functions of exponential, logarithmic, or hyperbolic types, there is a strong need for a better understanding of the reasons for this divergence phenomenon and for methods to overcome it and ensure convergence.

LEAST SQUARES METHOD

Let $\epsilon(X, t)$ be the discrepancy at point t between the experimental value $y_e(t)$ and the value of the approximation function $y(X, t)$. The method aims to minimize the least squares criterion function,

$$\phi(X) = \int [\epsilon(X, t)]^2 \, dt = \int [y_e(t) - y(X, t)]^2 \, dt, \tag{1}$$

with respect to the vector of parameters X (integral signs being taken in the Stiljes sense), or to solve the system

$$\delta\phi/\delta x_j = 0, \qquad \forall j, \tag{2}$$

which is equivalent to (1) in any domain where ϕ is unimodal.

First approximation. System (2) is expanded to the first order in the Taylor sense to give

$$-2 \int \left[\epsilon(X, t) - \sum_i \frac{\delta y}{\delta x_i}(X, t)\, dx_i\right]\left[\frac{\delta y}{\delta x_j}(X, t) + \sum_i \frac{\delta^2 y}{\delta x_i \, \delta x_j}(X, t)\, dx_i\right] dt = 0, \quad \forall_j . \tag{3}$$

Second approximation. In (3) we set

$$(\delta^2 y/\delta x_i \, \delta x_j)(X, t) = 0, \qquad \forall i, j .$$

This second approximation is equivalent to the following statement. y is approximated in the neighborhood of $y(X, t)$ by a function linear with respect to the set of x_j (cf. Geometric Properties), which has the same first derivatives at the point (X, t).

The system then takes the form

$$\sum_i \left[\int \frac{\delta y}{\delta x_j}(X, t)\frac{\delta y}{\delta x_i}(X, t)\,dt\right] dx_i = \int \epsilon(X, t)\frac{\delta y}{\delta x_j}(X, t)\,dt, \qquad \forall_j, \tag{4}$$

or, in a matric form, $B\,dX = E$, with

$$b_{ij} = \int \frac{\delta y}{\delta x_j}(X, t)\frac{\delta y}{\delta x_i}(X, t)\,dt; \qquad e_j = \int \epsilon(X, t)\frac{\delta y}{\delta x_j}(X, t)\,dt. \tag{5}$$

Note. When a discrete summation is used, it is customary to define

$$a_{kj} = (\delta y/\delta x_j)(X, t_k); \tag{6}$$

then $B = A^{\mathrm{T}}A$ and $E = A^{\mathrm{T}}\epsilon$, which provides a simple means to compute B and E.

Geometric Properties

We define

$$\phi = \int \epsilon^2\,dt; \tag{1}$$

thus

$$\begin{aligned}\frac{\delta\phi}{\delta x_i} &= 2\int \epsilon\frac{\delta\epsilon}{\delta x_i}\,dt,\\ \frac{\delta^2\phi}{\delta x_i\,\delta x_j} &= 2\int \frac{\delta\epsilon}{\delta x_i}\frac{\delta\epsilon}{\delta x_j}\,dt + 2\int \epsilon\frac{\delta^2\epsilon}{\delta x_i\,\delta x_j}\,dt.\end{aligned} \tag{7}$$

The resolution of $BX = E$ is equivalent to the minimization of the quadratic form

$$F(X) = \tfrac{1}{2}X^{\mathrm{T}}BX - X^{\mathrm{T}}E, \tag{8}$$

the derivatives of which are, at the point $X = 0$,

$$\begin{aligned}\frac{\delta F}{\delta x_i} &= (BX - E)_i\Big|_{X=0} = -\int \epsilon\frac{\delta y}{\delta x_i}\,dt = \int \epsilon\frac{\delta\epsilon}{\delta x_i}\,dt,\\ \frac{\delta^2 F}{\delta x_i\,\delta x_j} &= b_{ij} = \int \frac{\delta y}{\delta x_i}\frac{\delta y}{\delta x_j}\,dt = \int \frac{\delta\epsilon}{\delta x_i}\frac{\delta\epsilon}{\delta x_j}\,dt.\end{aligned} \tag{9}$$

The geometric interpretation follows by comparing $F(X)$ and $\frac{1}{2}[\phi(X) - \phi(0)]$; the surface ϕ, known only in $X = 0$, is replaced by a paraboloid $2F + \phi(0)$ with the same tangent plane and the same curvature in any plane containing $X = 0$; the minimum of the paraboloid is them searched for (Fig. 1).

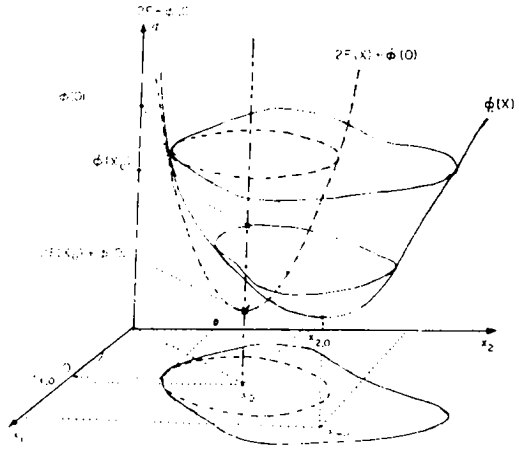

FIG. 1. Surface ϕ, the minimum of which is shown by solid lines. The approximation paraboloid is the dashed line. The figure shown $X_{\min}$, the true minimum; X_0, the computed X; and $\phi(X_0)$, the corresponding ϕ. The starting point of the iteration is (0, 0) with the ϕ_0 value for the criterion function.

PROPERTIES OF THE B MATRIX

(1) B is a real symmetric matrix. This clearly results from Eq. (5),

$$b_{ij} = b_{ji}\,.$$

(2) Diagonal elements of B are positive or zero (Eq. (5)):

$$b_{ii} = \int (\delta y/\delta x_i)^2\, dt.$$

(3) B is nonnegative definite.

Let X_0 be a solution of $F(X)$ minimum and $X = X_0 + V$; then

$$\begin{aligned} F(X) &= \tfrac{1}{2}[X_0 + V]^{\mathrm{T}} B[X_0 + V] - [X_0 + V]^{\mathrm{T}} E, \\ &= \tfrac{1}{2}X_0^{\mathrm{T}}BX_0 + \tfrac{1}{2}V^{\mathrm{T}}BX_0 + \tfrac{1}{2}X_0^{\mathrm{T}}BV + \tfrac{1}{2}V^{\mathrm{T}}BV - X_0^{\mathrm{T}}E - V^{\mathrm{T}}E, \\ &= F(X_0) + \tfrac{1}{2}(X_0^{\mathrm{T}}BV + V^{\mathrm{T}}BX_0) + \tfrac{1}{2}V^{\mathrm{T}}BV - V^{\mathrm{T}}E. \end{aligned}$$

Since B is symmetric, $X_0^T BV = V^T BX_0$, so that

$$F(X) = F(X_0) + \tfrac{1}{2}V^T BV + V^T(BX_0 - E),$$

and $BX_0 - E = 0$ since X_0 is a solution of $F(X)$ minimum; so

$$V^T BV = 2[F(X) - F(X_0)] \geqslant 0,$$

since $F(X_0)$ minimizes $F(X)$. Moreover, B can always be written as

$$B = T^T ST, \tag{10}$$

where T is the orthogonal matrix of eigenvectors and S is the diagonal matrix of the eigenvalues s_i (all s_i's being real, since B is real symmetric). If there is a vector $V \neq 0$ such that $V^T BV = 0$, then $(TV)^T S(TV) = U^T SU = \sum_i u_i^2 s_i = 0$, which shows that at least one s_i is zero. Then any vector $X = X_0 + kV$ is a solution of the system and $F(X)$ has another parabolic direction in the X space.

Minimization Methods of Levenberg and Marquardt–Meiron [1, 2, 3]

The aim of these methods is to modify the quadratic form F to bring its minimum nearer of the minimum of ϕ, while keeping, as much as possible, the properties of B. They use the fact that at least the gradient is known to be a direction where ϕ is decreasing.

These modifications are ($\lambda \geqslant 0$):

$$B_\lambda = B + \lambda I \qquad \text{(Levenberg)}, \tag{11}$$

$$\text{Diag}(B)^{-1/2}\, B_\lambda\, \text{Diag}(B)^{-1/2} = \text{Diag}(B)^{-1/2}\, B\, \text{Diag}(B)^{-1/2} + \lambda I \quad \text{(Marquardt)}, \tag{12}$$

$$B_\lambda = B + \lambda\, \text{Diag}(B) \qquad \text{(Meiron)}. \tag{13}$$

It is clear that

(1) Meiron's and Marquardt's transformations are identical;

(2) Marquardt's transformation is a Levenberg's transformation in a parameter space normed by $\text{Diag(B)}^{-1/2}$ (if $b_{ii} \neq 0, \forall i$).

So we can systematically reduce the study to that of Levenberg's method whenever all b_{ii}'s are nonzero.

Properties of the Matrix B_λ

Since the matrix B can be written as $T^T ST$ where T is the orthogonal matrix of eigenvectors, Eq. (11) becomes

$$B_\lambda = T^T ST + \lambda I = T^T ST + \lambda T^T T = T^T(S + \lambda I)\, T. \tag{14}$$

Hence, the matrix B_λ has the same eigenbase as B and its eigenvalues are $s_i + \lambda$; but the s_i values are positive or zero, since the quadratic form associated with B is nonnegative definite; the eigenvalues of B_λ are hence strictly positive for $\lambda > 0$, and the associated quadratic form is positive definite.

Cases of Singularity for B and Consequences in the Various Methods

We shall call "local" singularities those occurring for particular choices of the vector of parameters X, and "intrinsic" those.occurring for any X.

The singularities of B put a stop to the standard least squares method; two types of singularities can be distinguished.

(1) There is a subspace P of the parameter space X such that

$$\sum_{x_j \in P} \mu_j (\delta y/\delta x_j)(X, t) = 0, \quad \forall t \qquad \text{with} \quad \int [(\delta y/\delta x_j)(X, t)]^2 \, dt \neq 0, \quad \forall j.$$

Then, all the parameters of the subspace P are not simultaneously discernible; or, in other words, there is an infinity of vectors X, solutions of the minimization.

If the singularity is local or intrinsic, Levenberg's or Marquardt–Meiron's methods still lead to a solution since the linear dependence of the lines of the matrix B is destroyed in B_λ. If the singularity is a local one, the normal convergence process will start again when X has left the locus of singularities; if the singularity is intrinsic, at least one parameter is not independent of the others; the computed solution will minimize ϕ but will not be the only one because of the ill choice of parameters.

(2) There is at least one parameter x_k such that

$$(\delta y/\delta x_k)(X, t) = 0, \qquad \forall t;$$

then y is independent of x_k for the choice of parameters X. In this case, we have to study the methods separately, since Levenberg's method adds a λ to the diagonal term and leaves no singularity, Meiron's method leaves the kth line equal to zero, and there is no possible normalization in Marquardt's method since

$$b_{kk} = \int [(\delta y/\delta x_k)(X, t)]^2 \, dt = 0.$$

Yet, it must be considered that

if the singularity is intrinsic, y is independent of x_k, which makes this parameter meaningless;

if the singularity is local, a minimum in x_k is reached. It seems consistent not to change x_k, and since $b_{ik} = 0$, $\forall i$ and $e_k = 0$, one just needs to set $b_{kk} = 1$. The kth equation of (V) is $dx_k = 0$; the other equations are independent of dx_k since $b_{ki} = 0$ and can be solved if there is no other such singularity; otherwise, the same process will be repeated.

Note. Even very simple approximation functions y may exhibit such singularities, which make their linearized approximation from the first derivatives give a singular B matrix. For instance, let us consider

$$\begin{aligned} y &= x_1 e^{x_2 t} + x_3, \\ \delta y/\delta x_1 &= e^{x_2 t}, \\ -\delta y/\delta x_2 &= x_1 t e^{x_2 t}, \\ \delta y/\delta x_3 &= 1. \end{aligned}$$

Let $x_2 = 0$; then $\delta y/\delta x_1 - \delta y/\delta x_3 = 0$, a singularity of the first type.

Let $x_1 = 0$; then $\delta y/\delta x_2 = 0$, a singularity of the second type.

Though the y function is quite elsewhere, whenever the result of an iteration step leads to a solution in a vicinity of $X_1 = 0$ or $X_2 = 0$, the standard least squares method meets a nearly singular matrix (thus a very ill-conditioned system for inversion), and gives a poor convergence or a divergence in the iteration process.

Hence, the standard least squares method must not be used for nonlinear problems without a careful examination of the possible errors arising from local incompatibility between the approximation function y and the basic assumptions of the method.

Locus of The Solutions in the X Space

The hyperquadratics of the X space defined by

$$F_\lambda(X) = \tfrac{1}{2} X^T B_\lambda X - X^T E = 0$$

form a linear punctual sheaf, the basic hyperquadratics of which are

$$\begin{aligned} \tfrac{1}{2} X^T B X - X^T E &= 0, \\ \tfrac{1}{2} X^T I X &= 0. \end{aligned}$$

The solutions of the minimization of $F_\lambda(X)$ are the centers of the hyperquadratics. Their locus has, for asymptotic directions, those centers which make the poly-

nomial of the highest degree terls equal to zero. The locus is also the axis of the hyperparaboloids in the sheaf, defined by the nullity of the determinant associated to B_λ, which is an algebraic equation of degree n (equal to the number of parameters), and with all its roots real, $\lambda = -s_i$ (cf. properties of the matrix B). Hence they are the directions of the eigenvectors of B. These eigenvectors form an orthogonal base of the normed space of X vectors.

Therefore, the locus of the centers is a skewed algebraic curve of degree n, with n real asymptotes. Since the s_i's are positive or zero, all the solutions for $\lambda \geqslant 0$ are on a same continuous branch of the curve.

In a case with two parameters, hyperquadratics reduce to conics and the locus of the centers is the conic "of the nine points"; since the asymptotes are real, it is a hyperbola (Fig. 2).

In the same way, with three parameters, the locus of the centers of quadratics belonging to a linear sheaf is a skewed cubic curve.

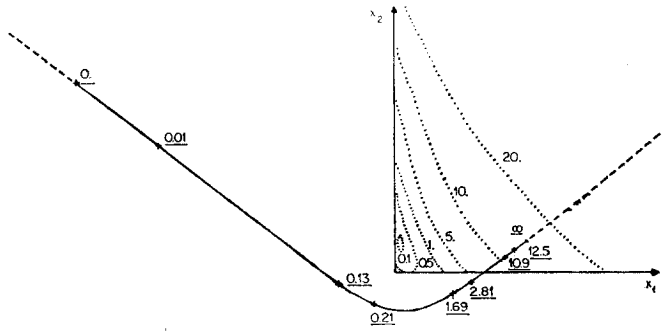

FIG. 2. Minimization on two parameters (heptane–ethanol mixture from [5]) showing the hyperbola locus of the solutions. Dotted lines are iso-ϕ. The function is not defined for X_1 or x_2 negative. Underlined values are λ's.

SELECTION OF THE OPTIMAL λ VALUE

Marquardt uses the smallest value of λ giving a convergence; this leads to the largest iteration step, but not necessarily to the smallest ϕ_λ. It seems more advisable to select the value of λ leading to the minimum of ϕ_λ. To do this, the general shape of the curves $\phi_\lambda(1/\lambda)$ must be considered. It is to be observed that ϕ_∞ corresponding to an infinite λ is known to be the result of the previous iteration, and ϕ_0 is the value obtained for $\lambda = 0$ by the standard least squares iteration. In the simple case of a unimodal ϕ_λ function for $0 \leqslant \lambda$, three types of curves can occur (Fig. 3):

(3e) There is no minimum and the standard method is optimal;

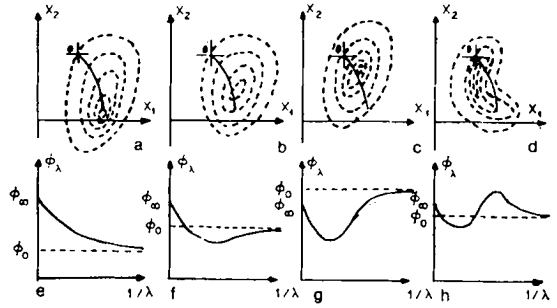

FIG. 3. (a, b, c, d): iso-ϕ curves in a two-parameter space and locus of X_λ $(0 \leqslant \lambda)$; 0 is the origin of the minimization corresponding to λ infinite and $X_\lambda = 0$. (e, f, g, h): corresponding ϕ_λ $(1/\lambda)$ curves. (a, c): standard least squares are optimal; (b, f): standard least squares converge, nonoptimal; (c, g): standard least squares diverge; (d, h): complex case where $\phi_\lambda(1/\lambda)$ is not unimodal.

(3f) there is a minimum and an optimal λ value, though the standard least squares still converge since $\phi_\infty > \phi_0$;

(3g) there is a minimum and an optimal λ value; since $\phi_\infty < \phi_0$ the standard least squares diverge.

If ϕ_λ is not unimodal, there is an optimal (zero or nonzero) λ, but the search will be much more complex.

One method to find the optimal λ is to approximate ϕ_λ by a function with p parameters, the minimum of which can be computed readily enough.

Two parameters can be determined by the conditions for λ infinite.

(1) $\phi = \phi_\infty$, which is known;

(2) $[d\phi/d(1/\lambda)_\infty]$ is readily computable since the initial system, Eqs. (11) and (5), gives

$$(B + \lambda I)\, X = E \qquad \text{or} \qquad X = (B + \lambda I)^{-1} E,$$

and derivation with respect to $1/\lambda$ leads to

$$-\lambda^2 X + (B + \lambda I)(dX/d(1/\lambda)) = 0 \qquad \text{whence} \qquad dX/d(1/\lambda) = \lambda^2 (B + \lambda I)^{-2} E.$$

The value of ϕ for a vector X close enough to zero is equal to that of $2F + \phi_0$, that is,

$$\phi = X^T B X - 2X^T E + \phi_0,$$

whence

$$\begin{aligned} d\phi/d(1/\lambda) &= 2X^T B(dX/d(1/\lambda)) - 2E^T(dX/d(1/\lambda)), \\ &= 2\lambda^2 E^T (B + \lambda I)^{-1} B(B + \lambda I)^{-2} E - 2\lambda^2 E^T (B + \lambda I)^{-2} E. \end{aligned}$$

For λ increasing to infinity, $(B + \lambda I)$ tends toward λI and $(B + \lambda I)^{-1}$ toward $(1/\lambda)\, I$, so the first term is of $1/\lambda$ order and tends to zero; the second term has a limit equalt to $-2E^TE$. Thence

$$[d\phi/d(1/\lambda)]_{-\infty} = -2E^TE. \tag{16}$$

The remaining $p - 2$ parameters have to be determined from ϕ_λ values computed for selected λ (such as the optimal λ of the previous iteration) [4].

Conclusion: Numerical Experience

These methods were applied to problems where the standard least squares method was very slowly convergent or failed to converge, such as the determination of interaction coefficients in liquid binary mixtures [5], the fitting of infrared absorption hand envelopes, or models for chromatographic peaks on experimental results [6].

The Marquardt–Meiron method always exhibited a very good convergence (often superior to that of Levenberg's) in the most difficult cases, where the quadratic approximation of the least squares criterion function was very poor due to the high nonlinearity of the y functions used, even for an initial guess of the parameters far from the true value.

References

1. K. Levenberg, *Quart. Appl. Math.* **2** (1944), 164.
2. D. W. Marquardt, *J. Soc. Ind. Appl. Math.* **11** (1963), 431.
3. J. Meiron, *J. Opt. Soc. Amer.* **55** (1965), 1105.
4. J. Pitha and R. N. Jones, *Canad. J. Chem.* **44** (1966), 3031.
5. B. Aubineau, Thése de Docteur-Ingénieur, Université de Paris VI, Paris, 1973, CNRS A.O. 8728.
6. F. Haffner and J. P. Petit, International Congress, "Use of Electronic Computers in Chemical Engineering," Paris, 1973.

Received: December 10, 1974; revised: May 17, 1976

Henry Brusset
Dominique Depeyre
Jean-Pierre Petit
Francois Haffner

Laboratoire de Génie et Informatique Chimiques
Institut de Chimie
École Centrale des Arts et Manufactures
92290 Chatenay-Malabry, France